A Philosophical Interpretation of Judea Pearl's Theory of Causality

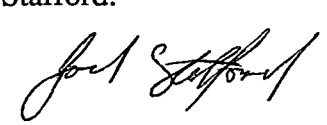
by
Joel Stafford, BA(Hons)

Submitted in fulfilment of the requirements for the degree of
Master of Arts
University of Tasmania
August, 2004.

Candidate Statement

I the undersigned declare that this thesis contains no material which has been accepted for a degree or diploma by the university or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgment is made in the text of the thesis.

Mr. Joel Stafford.



Acknowledgements:

Numerous people helped me to avoid errors in the preparation of this thesis. I own all those remaining, several of which I am aware and can do nothing about, having reached the limits of my abilities. Thanks are due to Emily Hansen, the Hansen family, the Stafford family, Toby Patterson, Mark Bravington, Phil Dowe, Stephen Barker, David Coady, James Chase, Hannah Jenkins, Graham Wood, Mitch Parsell, Bruce Tranter, Everarda Cunningham, Brian Graetz.

Authority of Access

Abstract:

Judea Pearl has provided a book long treatment to the topic of causality from a formal and predominantly non-philosophical perspective. One of the main aims of Pearl's treatment is to increase the clarity of scientists' reasoning about problems involving causal relationships. I offer a philosophical interpretation of Judea Pearl's treatment of causality. I argue Pearl (2000a) articulates an analysis of causality in two distinct parts. I claim that Pearl unifies several of the key concepts of causality relevant to the epidemiological, economic, social and biological sciences on one part and that Pearl describes several general conditions characteristic of objective causal processes on the other. I label the first part Pearl's account of causal modelling and the second part Pearl's account of causal processes respectively. I find that Pearl's account of causal processes is of the greatest philosophical interest despite the fact that Pearl's account of causal modelling is the primary component of the overall treatment. I locate Pearl's account of causal modelling first within the recent philosophical literature on the nature of models and modelling practices before the literature on causation and explanation as has been the standard elsewhere. With some reservations I argue Pearl's views on objective constraints are consistent with a nomic view of causation. To focus my interpretation I discuss several philosophical criticisms of Pearl's account and argue that each one is misplaced when targeted at the first part of the account and merely analogues to familiar problems of model interpretation and justification when targeted at the second. I conclude that Pearl's account provides good reason to re-sharpen the philosophical focus on what it is to give an analysis of causation.

A Philosophical Interpretation of
Judea Pearl's Theory of Causality

"Insofar as geometry is about experience it is not certain, and insofar as geometry is certain it is not about experience"

Albert Einstein.

1.0 Introduction

Some time ago Vaughn McKim asserted that the goals of non-experimental research in the sciences required an urgent and fundamental rethinking. McKim implied that methods designed to draw causal conclusions in these sciences were in a state of crisis (McKim and Turner 1997:19). Today many hold the view that little has changed. What has changed is where the blame is laid. It is our definition of causality, which has been rethought, and not the belief that causal knowledge is achievable in non-experimental research. In the time since McKim's comments about a crisis in causality the goals of causal modelling in non-experimental settings have received several formulations and, principally through the work of Pearl (2000a), have undergone a fundamental rethinking. At present the theoretical work behind causal modelling is as much about shifting causal reasoning beyond the realm of statistical analysis as it is about aiding in the discovery of causal relationships. Contemporary causal modellers in sciences such as biology and epidemiology look upon the suite of distinct and apparently incompatible theories of causation painstakingly built by philosophers, not as theories competing to articulate the correct account, but as a toolkit to be exploited and adapted to solve problems on a case-by-case basis.

Contemporary causal modelling methods presuppose that the world is a world of causation rather than mere correlation and that the patterns exhibited in quantitative descriptions of the world offer a window through which to identify causal relations

when few or no experimental methods are applicable. This is the view that causal relationships produce statistical relationships. According to this view, non-experimental techniques are used in the social and behavioural sciences, epidemiology and biology to test hypothesised causal relations underlying sets of observational data. In each of these fields a premium is placed upon the identification of mechanisms and it is becoming increasingly common to look to modern modelling techniques to build mathematical structures toward their identification and description. The general goal, held in common with experimental techniques, is that of contributing to the scientific understanding of reality through discovery. A subtle but seemingly indispensable part of the project of discovery using non-experimental methods is the construction of a language purpose built for the clear expression and evaluation of causal claims.

Over the last decade two prominent and distinct causal modelling research groups have emerged, both aiming to achieve the goal of clarifying causal language and inference in science. Spirtes, Glymour and Scheines (1993) and the TETRAD program represent the centre of one group and Pearl (1988, 2000a) represent the centre of the other. Both programs aim to design algorithms capable of discovering and quantifying causal relationships in data sets by searching and locating in data some number of equivalent 'causal structures' consistent with the probabilistic dependence and independence relationships the data exhibit. The standard view of these projects is that they have two primary aims. The first is inference. Causal modelling programs aim to establish formal methods for drawing causal conclusions from statistical data. The second is clarification. Causal modelling programs aim to clarify the role that causality plays in science in order to positively affect the way in which practicing scientists construct models and collect data. However, according to the standard view, causal modelling programs discuss and work with 'causal' concepts toward the realisation of these aims without first providing a definition of causality. Instead, causal modelling programs 'axiomatize' familiar causal notions within a formal language and then use this language to express, quantify, and test causal claims. It is in this sense, it is thought, that causal statements do not report the

observations of the investigator, but instead are conditional on the axioms and licensed by the inference rules of a formal calculus. As Pearl admits, causal modelling programs attempt a 'mathematization' of causality.

The causal notion of greatest importance to the mathematization of causality is expressed through the Causal Markov condition. According to the Causal Markov condition the value of a variable in a causal model is independent of that variable's effects, when conditioned on its direct causes. Stating that a model is causally Markov means that, ignoring the effects of any given variable in the model, all relevant probabilistic information about the model's variables may be discerned from each variable's direct causes. The underlying notion is that the state of a system at one given moment is relevant to its subsequent state in a way that the prior states of the same system are not (Scheines 1997:190). Concentrating upon a variable and its direct causes in this way lends a causal model a modular character and invites consideration of how a model might respond if its individual modules were to be altered. At this point it seems natural to compare the alteration of a model's components with experimental manipulation and it is commonplace that comparison between experimental and non-experimental methods is described in terms where the latter becomes a simulation of the former. The common ground between the two is that in both cases the investigator wishes to gain knowledge of causal relationships. The key difference between the two is that instead of identifying and quantifying cause/effect relationships by allowing nature to run its course within a suitably monitored and isolated space, as with experimental methods, non-experimental model manipulations follow the rules of a formal calculus. According to Michael Freedman, it is here that causal modelling programs violate the *law of conservation of rabbits: If you want to pull a rabbit out of the hat, you have to put a rabbit into the hat* (Freedman 1997: 181-182). In other words, to discover causal relationships using non-experimental methods you have to put them in to the model in the first place, and this defies elementary principles of science.

I deny that Pearl's theory of causality is at odds with elementary principles of science. As a first step towards a clear demonstration of this claim I aim to set out in this thesis a philosophical interpretation of Pearl's theory of causality. Second, there is, to the best of my knowledge, no widespread agreement on the correct classification of Pearl's theory within the body of philosophical work on causation, nor is there any consensus regarding the theory's fundamental features. I intend that my interpretation will specify Pearl's theory's fundamental features and correctly locate the theory in the body of philosophical work on causation.

I present the thesis over three chapters. The interpretation I develop is that Pearl's theory of causality consists of two distinct components or 'tiers'. I label one of the two 'Pearl's regimentation' and explain how it is an account of causal modelling in science. I label the other tier 'Pearl's objective account' and argue that it amounts to a partial analysis of causal relations as constraints on physical processes. In the first chapter I set the ground for the presentation of Pearl's theory and the interpretation to follow. I identify in chapter 1 that Pearl's theory has important connections with the so-called causal/mechanical tradition in philosophy but that Pearl's theory is by-and-large discontinuous with the philosophy of science on issues involving explanation and causation. I find that an adequate philosophical interpretation must connect the preoccupations of the philosophy of science and the preoccupations of scientific modellers. To make the connection I look to several emerging projects in the philosophy of science, which focus on the nature of scientific models. I identify the account of mechanisms recently offered by Stuart Glennan as a natural philosophical counterpart for Pearl's concept of causal structure. Last, I compare Pearl's orientation towards causation with the orientation attributed to him by contemporary philosophers of causation and explanation. I suggest that the case for classifying Pearl as a manipulation theorist is not clear-cut.

In chapter 2 I articulate what I see as the primary components of Pearl's theory of causality in his (2000a). I describe these components in two sections, the first presents the syntactic features of the theory and the second presents its semantics.

The presentation traces a path that begins with the mathematical theory of graphs, especially Directed Acyclic Graphs, moves through the detail of Pearl's logic of causal reasoning and ends with a discussion of how causal expressions of the logic may be deduced from a causal model. Following my presentation of Pearl's formalism I offer an informal interpretation. The most important components I discuss include Pearl's interpretation of his formalism's counterfactual semantics; the theory's account of types versus token causes; and the theory's interpretation of mechanisms. In the final section of chapter 2 I discuss the fact that Pearl's theory is committed to a boundary between statistics and causality. The discussion centres on Pearl's view that statistical modelling is limited to the estimation and manipulation of expressions that represent 'static' observations, whereas causal models extend to the representation of mechanisms responsible for the production of observations. I argue it is a puzzle how Pearl is able to move from a statistical interpretation of the relationships within a Bayesian Network to a causal interpretation of the same network when the former encodes degrees of belief and the latter encodes mechanisms representing objective law-like relations in the natural world. Finally, after suggesting how the puzzle might be resolved I explain that the subjective and objective components identified in the preceding discussion are suggestive of two distinct uses of causality in Pearl's theory.

In chapter 3 I argue that the two distinct senses of causality identified in the previous chapter are reflective of two distinct components to Pearl's theory of causality. I claim here that Pearl's theory of causality is the conjunction of an account of causal modelling in the applied sciences and an account of the natural properties assumed to exist by those sciences. I call the first conjunct 'Pearl's regimentation of causal concepts' and claim it resembles a conceptual analysis but really isn't one, and I call the second conjunct 'Pearl's objective account' and claim it resembles an incomplete empirical analysis of the causal relation. I then proceed to set out my view that Pearl's account of causal modelling involves the explicit specification and regimentation of several important causal concepts and a set of procedures for their application in non-experimental sciences. Then I discuss Pearl's view that the world

consists of numerous autonomous invariant linkages that correspond to physical processes and compare this view to what philosophers have called nomic causation. After detailing my interpretation of what Pearl is saying about causality I look to deepen the discussion by exploring a number of criticisms levelled against Pearl's theory. I argue that several key criticisms of Pearl's theory are misplaced because each mistakenly assumes that Pearl's theory is nothing over-and-above what is on my interpretation Pearl's regimentation. In discussion I identify that each criticism in effect amounts to a challenge of a model's credentials to represent causal relationships. In concluding this chapter I suggest that credentialing a causal model is not an all or nothing affair and draw a link between this issue and contemporary philosophical literature concerned with ontological commitment.

1.1 Causal Modelling in Science

I begin this chapter with an introduction to issues pertinent to causal modelling in science. But, before I attend to those details I will need to say something about why I begin with a discussion of causal modelling when, as I claim in the introduction, one of the key aims of this research is to offer an interpretation of Judea Pearl's theory of causality. The answer is straightforward. Pearl (2000a) is as much about procedures for building and vetting models of causal interactions in various areas of science as it is about the nature of those causal interactions. However, there is a twist here involving the relationship between model building procedures and analysing causation that will take a moment to sketch. The onus of adequately interpreting Pearl's (or anyone's) theory (of causation) rests, of course, with the interpreter. Interpreting a causal theorist's intentions arguably brings with it more pressing concerns than those to be found in other areas of the philosophy of science due to the fact that causal concepts are pervasive throughout natural language and are wrapped up in everyday human practice. Causal concepts can be difficult to isolate. Attempts to draw technical definitions of familiar causal terms can often add more confusion than clarity, especially when theorists draw on intuitions or make use of examples and analogies pertaining to the everyday world. In short, the task of interpreting and analysing a theory of causation is probably second in difficulty only to that of constructing a workable theory of causation itself mainly because it is especially difficult to analyse something without the benefit of sufficient distance.

There are numerous ways to discuss the philosophical study of causation. Granted that causation involves some form of relation, philosophical interest in causation may be summarised as the attempt to answer the following sorts of questions. Beginning with what is perhaps the question of greatest importance, philosophers interested in causation wish to know what distinguishes causal from non-causal sequences.

Besides this question philosophers wish to know what are the relata of the causal sequences, how many relata there are to the relation, how are these relata individuated, and how to explain the apparent asymmetry of the causal relation. Some have argued that the category of causation may be eliminated so far as science is concerned (eg. Russell 1913), and others have claimed that causation must be taken as a primitive notion (eg. Anscombe 1993, Tooley 1993). Those who accept that causation exists and is amenable to analysis have produced numerous distinct theories. For instance, probability theories claim that the causal relation must involve a probability increase or change (eg. Suppes 1970; Cartwright 1979; Eells 1991), process theories claim that causation must involve, for example, the world lines of objects that possess a universally conserved quantity (eg. Skyrms 1980; Dowe 2000), or property transference (eg. Aronson 1971; Ehring 1998). Counterfactual theories claim that the causal relation must involve some form of counterfactual dependence (eg. Lewis 1973b, 1986; Ramachandran 1997), and other theories offer hybrid analyses (eg. Schaffer 2001).

Philosopher's interested in theorising causation expend considerable effort trying to situate their theory amongst existing projects. As is the case with theories of causation, there are several ways to describe contemporary projects involved with the philosophy of causation. One division categorises a project according to whether its aim is to provide an analysis of what causation is in contrast to an analysis of what natural language causal utterances mean and how (folk and specialised) causal concepts are best mapped (Sosa and Tooley 1993; Armstrong 1997; Jackson 1994; Bigelow and Pargetter 1990). An alternative approach groups projects according to their acceptance or denial of key Humean ideals about causation and laws of nature, such as the so-called Humean supervenience thesis (eg. Psillos 2002)[1]. What is noteworthy is that no matter the approach adopted, philosophers tend to agree the task

[1] These issues are recounted in part across a large literature including, for instance, Beauchamp and Rosenberg (1981), Mackie (1974), Sosa and Tooley (1993), Lewis (1986), Mellor (1995), Salmon (1984, 1990, 1998), Sankey (1999), and Psillos (2002) to name a few.

of constructing and assessing causal theories is one best carried out by philosophers[2] Likewise, analysing what causation *is* has typically been taken throughout the literature to be (at least primarily) a task built for philosophy[3].

Here enters the twist. In my reading of Pearl (2000a) I discover his project cuts across those approaches and attitudes taken by the majority of philosophers of science when analysing (or constructing) an account of causation. I discover that, taken broadly, Pearl's present approach is oriented towards the needs and requirements of the scientific modelling community[4] and although attentive to some philosophical issues, the practices, outcomes, and measures of success tend to differ in that area compared to those which are cited throughout, and which to some extent propel, key projects in the philosophy of science[5].

Modellers in science are charged with a task made difficult by numerous factors. These include such things as lack of suitable formal tools and computational power, time constraints, poor data, lack of adequate funding and expertise, lack of research 'breakthroughs' or successes, and in many cases just the brute intellectual complexity of attempting to comprehend and represent the phenomena under investigation with the myriad tools and techniques available. With such factors conspiring together, the task of the modeller is liable to be driven more by the need to achieve a meaningful conclusion, no matter how partial, than by following received doctrine or by being attentive to so-called 'foundation issues'[6].

---

[2] That is, within the body of philosophical work relevant to causation. See, for instance, comments by Psillos (2002: 3).

[3] This can be read as both a compliment and a slight. Those who claim that causation is mysterious and metaphysical slight philosophers when they claim it is the philosopher's task to analyse it.

[4] Exactly what community this term refers to is difficult to specify precisely. Briefly stated, I have in mind working scientists engaged in non-experimental techniques of data analysis. This research has a large following spanning across several disciplines. I detail the orientation of some groups within this community below.

[5] For instance, Menzies (2002) comments that Pearl's (2000) and (2001) have unfortunately not been broadly acknowledged by causal theorists let alone the philosophy of science community. See also similar comments by Gillies (2001).

Pearl's program is attentive to and appears to have developed in parallel with the difficulties encountered 'on the ground' by modellers far more so than the approach with which the philosophy of science has tended to tackle causal theorising. The reader will appreciate then that there is a gulf to be bridged between Pearl's approach to causality and those approaches commonly adopted in the philosophy of science before any interpretation can commence. There are several reasons why the gulf exists. One is simply because philosophers rarely initiate or participate in scientific investigations first hand and so are somewhat insulated from day-to-day pragmatic issues, problems, and developments. But this aspect of the gulf is not what I have in mind and so can be noted and set aside. What I see as the primary reason behind the existence of the gulf will take some explaining. A good place to start is with scientific theories and models.

There are numerous philosophical issues surrounding the task of building models in science. Many of these are related to broader issues in the philosophy of science including how to account for the structure of scientific theories; how to account for progress and discovery in science; questions that concern the nature of scientific methodology and the role of explanation and of truth; appropriate ways of confirming hypotheses; and questions related to analogical reasoning and how models mediate between theories and the world, to name but a few. Building models that aim to capture causal relations or that licence causal inferences has impacted upon each of these areas of investigation and their associated difficulties and impasses. Indeed calling them 'issues' in the philosophy of science is to understate their importance. At one time, and for many philosophers of science today, such issues are the central concern. For instance, Frederick Suppe expresses the following sentiments concerning the constitution of the philosophy of science:

> If any problem in the philosophy of science justifiably can be claimed the most central or important, it is that of the nature and structure of scientific

[6] This is to claim not that scientists and modellers aren't attentive to foundational issues in their respective disciplines or don't conform to received doctrines but that these are typically the first to go

> theories, including the diverse roles theories play in the scientific enterprise. For theories are the vehicle of scientific knowledge, and one way or another become involved in most aspects of the scientific enterprise. It is only a slight exaggeration to claim that a philosophy of science is little more than an analysis of theories and their roles in the scientific enterprise. A philosophy of science's analysis of the nature of theories, including their roles in the growth of scientific knowledge, thus is its keystone; and should that analysis prove inadequate, that inadequacy is likely to extend to its account of the remaining aspects of the scientific enterprise and the knowledge it provides (Suppe 1977: 3)[7].

Woodward (2003) confirms and focuses these sentiments:

> Issues concerning scientific explanation have been a focus of philosophical attention from Pre-Socratic times through the modern period. However, recent discussion really begins with the development of the Deductive-Nomological (DN) model. This model has had many advocates (including Popper 1935, 1959,
> Braithwaite 1953, Gardiner 1959, Nagel 1961) but unquestionably the most detailed and influential statement is due to Carl Hempel (Hempel, 1942, 1965, Hempel and Oppenheim, 1948). These papers and the reaction to them have structured subsequent discussion concerning scientific explanation to an extraordinary degree (Woodward 2003: 1).

As does Psillos (2002) when he asserts that the pervasive nature of causal and explanatory talk in the sciences elevates its importance to a level that can hardly be exaggerated (Psillos 2002 p 1). Most would agree that to understand what science tells us about the way the world is requires analysis of the nature of scientific theories. Increasingly the analysis has tended to focus on explanation and causation.

---

when Nature fails to come to the party.

I agree with Psillos (2002) that the connection between causal and explanatory talk is important to highlight. For some time one of the primary areas for research into causation has been the project of accounting for scientific explanation. A major tenet of this project is that explanation is the main goal of scientific theorising. In Salmon's words:

> Science, the majority say, has at least two principal aims–prediction (construed broadly enough to include inference from the observed to the unobserved regardless of temporal relations) and explanation. The first of these provides knowledge of *what* happens; the second is supposed to furnish knowledge of *why* things happen as they do (Salmon 1978: 684, emphasis in original).

But if scientists hold the same sentiments as philosophers on such matters they rarely express them. Scientists involved with modelling have tended to treat causation (if at all) as a technical difficulty that, to be overcome, requires advances in the types of formalism and testing procedures that lead to greater predictive power, a simplification or unification of current methods, or the introduction of entirely new methods. As Pearl remarks, scientists need to track down cause-effect relations from the environment via limited actions and noisy observations (Pearl 2000a: 42-43). Analysing theory structure does not seem necessary to such practices. Moreover, the ability to identify and quantify causal relations is, according to Pearl (2000a), the ability to answer 'how' (and/or 'what-if') questions, such as: 'How to shape the beam so that it will carry the required load?' and 'What if the beam were narrower; would it still carry the load?' rather than 'why' questions (Pearl 2000a: 343)[8]. Pearl's

---

[7] Recently Suppe has rejected this view. See Suppe (2000), especially pp 109-110, and below for discussion.

[8] This is controversial. Pearl is open to the charge that what he hears as 'how' questions are really just Salmon's 'why' questions in disguise. One could further claim that Pearl's views are consistent with Karl Pearson's regarding the aims of science. There is some truth in this latter point. However, I should think Pearson would strongly disagree with Pearl's thesis on causality. It is worth noting that Salmon finished his presidential address by stating that that scientific explanation offers, over and above the inferential capacity of prediction and retrodiction, 'knowledge of the mechanisms of production and

comments imply that knowledge of how a system works comes prior to and is of greater importance than knowing why. The difference is partly due to Pearl's perspective of explanation in science. Pearl's perspective on explanation in science can be partially summarised by the following three assertions[9]:

> 1. Providing an account of scientific explanation requires a prior account of causality.
> 2. The principal task that besets an account of causality and explanation is providing procedures that aid the search for answers to 'how' and 'what-if' questions given a specific investigator-relative context.
> 3. Success in prediction and explanation follows from the ability to answer 'how' and 'what-if' questions and is parasitic on the ability to express causal queries in a clear language, devise successful 'experiments'–whether actual or hypothetical–and/or construct appropriate apparatus.

Pearl's view of explanation places the identification of mechanisms in a central position, and since, for Pearl (2000a), causation underpins explanation, the identification of mechanisms is the backbone of his causal discovery program. This means that modelling causal systems primarily involves articulating a procedure for representing what a system is doing and how it is doing it in such a way that one may know how the system would behave if it were altered. Pearl's view holds little resemblance to the project philosophers identify as explanation by unification and instead is reminiscent of the causal/mechanical tradition and what Salmon and others call the causal/mechanical project of explanation in science[10]. According to Salmon (1998), from the causal/mechanical view 'one looks at the world and its furniture as black boxes whose internal workings cannot always be directly observed but where

---

propagation of structure in the world' and that he held the view that 'knowledge of the mechanisms of production and propagation of structure in the world yields scientific understanding, and that this is what we seek when we pose explanation-seeking why questions'. In light of this one could argue that Salmon's 'why' questions are really Pearl's 'how' questions.

[9] See Pearl (2000a: 334-335) and Halpern and Pearl (2001b).

[10] The former project is primarily identified with the work of Kitcher and Freidman. For details and further discussion see Kitcher (1976, 1981, 1985, 1989), Freidman (1974), and Salmon (1985, 1990, 1997).

science's overriding aim is to open those boxes and expose the mechanisms inside' (Salmon 1998: 77). On this view scientific explanation aims to provide *understanding* where understanding results from knowing *how things work* (Salmon 1984: 240)[11]. Similarly for Pearl, understanding is the result of knowing how things respond to novel interventions and alterations (Pearl 2000a: 25-26).

However, despite these connections, I find that the resemblance is not enough on which to base an interpretation of Pearl (2000a). Pearl (2000a) diverges from the causal/mechanical tradition in several respects. First, when Salmon speaks of 'the world and its furniture' he means to make no distinction between the macro and micro world, whereas Pearl (2000a) intends only to take account of causal processes in the macro world[12]. Second, rather than focusing on the relationship between a theory and its models, as is typical in the causal/mechanical tradition, Pearl attempts to elicit how things work via appropriate modelling procedures focussed on getting a model to 'fit the data' conditional upon the context provided by the investigator's scientific specialty and present state of knowledge[13].

The divergence has several consequences. The most salient consequence is that it marginalizes the relevance of the philosophical literature on the structure of scientific theories and its accompanying account of scientific explanation. Hence, the divergence also displaces the importance of analysing and interpreting the components of Pearl's theory within the context of the semantic account of theories[14]. However, since Pearl aims to construct a formalism suitable for causal reasoning and explanation, formal models are not dispensed with altogether. Therefore, to both illustrate the extent of the divergence and to identify an acceptable context with

---

[11] Compare comments by (Pearl 2000: 334-336; 343; 345).

[12] And even then only from the perspective of specific disciplines.

[13] But where fitting data is carried out not just with statistical techniques but with causal techniques also. I discuss the difference in sections 2.0.3 and 2.1 of chapter 2. The requirement that explanation take place within an investigator relative context lends the account an epistemic dimension. I discuss the dimension in chapter 2 when I detail Pearl's account of counterfactual conditionals. Note also that Pearl tends to identify mechanism with structure. Of course, exactly what the identification amounts to forms one of the main preoccupations of this thesis.

[14] This divergence is discussed in some depth by Woodward (2000). But, see also Suppe (2000).

which to commence an interpretation it is now necessary to briefly discuss the semantic view of theories.

Scientific theories have tended to be analysed by philosophers of science as formal structures. It was once thought that understanding what a scientific theory tells us about the world must flow from the logical analysis of the theory's structure. Such logical analysis proceeded on the assumption that the linguistic expression of a theory could be reconstructed as an axiomatic system formulated in the language of mathematics and predicate logic. Understanding a specific theory meant drawing a connection between the syntactic features of the reconstructed theory and a given stock of observation terms via a set of correspondence rules. For several reasons this 'classical' view of theories was dispensed with. In the move away from the classical view of scientific theories, the 'semantic view' characterises scientific theories as sets of models conceived as mathematical structures. The difference between the two views is that, on the semantic view, theories are identified with a set of models, such that, in contradistinction to the classical view, the set remains the same no matter in what language it is expressed (Teller 2001: 394). According to Suppe (1974):

> Theories [on the semantic conception] are extralinguistic entities which can be described by their linguistic formulations. The propositions in a formulation of a theory thus provide true descriptions of the theory, and so the theory qualifies as a model for each of its formulations. This suggests that the semantic techniques of model theory [...] will be useful in analysing the structure of scientific theories. This suggestion gains further plausibility when it is noted that in actual practice the presentation of a scientific theory often takes the form of specifying an intended model [...] (Suppe 1974: 222 quoted by Glennan 2000a: 2)

Thus, the semantic conception includes the general claim that theories are semantic rather than syntactic entities and so the apparatus of formal semantics and model theory illuminate questions about the nature of scientific theories and models. On the

semantic conception the class of scientific models is thought of as a proper subset of the class of semantic models. In particular, the set of scientific models of a scientific theory is just the set of intended models of a formulation of that theory. Furthermore, on the semantic conception a scientific theory is just the class of intended models of one of the (equivalent) formulations of that theory. Hence, a scientific model is a form of semantic model and scientific theories are collections of semantic models, which apply to the world by the relation of isomorphism between models and some parts of the world (Glennan 2000a: 2; Teller 2001: 394).

The term 'semantic model' has been variously defined in terms of set theory and state spaces. The set-theoretic approach, due chiefly to Suppes (1960, 1967), takes scientific models to be semantic models familiar from presentations of the semantics of predicate logics. Giere (1999) points out that according to Suppes's thesis the meaning and use of models can be interpreted as being the same in the empirical sciences as it is in mathematics and mathematical logic. Glennan (2000a) provides the following summary:

> In such presentations [of the semantic conception], a model is defined as an interpretation of a set of statements of predicate logic under which all members of that set are true. An interpretation is in turn understood to be a function from non-logical symbols of the language onto individuals or sets of individuals of a given domain. If we consider the image of this function, we have a set-theoretic structure that is said to satisfy the set of statements [presenting the theory] (Glennan 2000a: 3).

Alternatively, on the state space approach, the state of a physical system is characterised by a set of variables that may measure the values of various physical magnitudes, such that the set of logically possible states of the system is to be identified with the set of all possible combinations of values of each of the variables. In turn, these combinations are treated as vectors in the state space and the dynamical behaviour of a modelled system may then be characterised in terms of the trajectory

of the system through this vector space over time. Laws of succession or coexistence may then be defined to characterise physically possible changes in the state of a system and physically possible combinations of values of state variables respectively (Suppe 1989)[15].

However, no matter which presentation is used, on the semantic conception models are emphasised in the main only in-as-much as they are taken to satisfy the theory's specifications. It is theories that are held to 'define a class of ideal systems which are then held (via theoretical hypotheses) to represent actual physical systems' and so models are important only in as much as it is important to be attentive to their identification with their respective theory (Glennan 2000a: 2, 5, 7). That is, to reiterate a point from above, scientific theories are on the semantic conception just the class of intended models of one of the equivalent formulations of that theory.

But, there are a number of ways in which models are thought to mediate between laws and physical systems. For instance, a model may specify the idealised conditions under which laws can be appropriately applied to a system, or, a model may specify how a combination of general laws or principles combine to apply to a particular case (Glennan 2000a: 6-7). The latter has the most in common with Pearl's view of the role models play in causal discovery because, for Pearl, models play the role of mediator; a modeller attempts to construct models that have some kind of predictive relationship to a complex natural phenomenon. Through clarifying the questions the investigator has about the causal relationships in the system under study, the investigator uses the model to solve causal problems. As such and in contrast with the semantic conception, the relationship between model and theory is not tight since, as is clear from practice, models are not hatched straightforwardly from theories:

> While models generally incorporate a great deal of the theory or theories with which they are connected, they are usually fashioned by appeal to, by inspiration from, and with the use of material from, an astonishingly large

[15] See also van Fraassen (1980; 1987).

> range of sources: empirical data, mechanical models, calculational techniques (from the exact to the outrageously inexact), metaphor, and intuition (Winsberg 2003: 106).

The result is that the applicability of the semantic conception is curtailed precisely because it identifies theories with models or classes of models but not, say, with theory fragments[16].

Adherence to the semantic conception's view of scientific models obscures the role that modellers perform when they attempt to build models from a collection of theory fragments or a collection of laws and principles[17]. The differing orientation of the modelling community to sections of the philosophy of science turns on the fact that philosophers have tended to gravitate towards explicating the relation between theories and their models whereas scientific modellers (including Pearl) are concentrating on exemplifying the relation between models and phenomena[18](Glennan 2000a). The consequence is this: the semantic view of scientific theories and the causal/mechanical account of explanation do not together nor separately make an adequate context within which to commence an interpretation of Pearl's account of causality. Evidently, Pearl disagrees with the spirit of Suppe's assertion that 'theories are the vehicle of scientific knowledge', since, for Pearl, focussing on theories obscures the role models and competition between models play in explanation. Pearl's project diverges from the causal/mechanical project in terms

---

[16] In fact, Suppe has recently come to reject his earlier characterisation of theories. See Suppe (2000) and Norton and Suppe (2001) for discussion.

[17] As a consequence, even a rational reconstruction of the investigator's model represented as a semantic model can be at odds with or have no obvious meaning given the intended interpretation of the collection of theories (fragments) employed by the investigator.

[18] With, of course, the aim of making correct predictions and solving actual problems. This divergence is particularly notable when one takes into account Pearl's attempt to define 'actual causation'. The point stands even in relation to recent works that question the semantic conception's account of scientific explanation such as Woodward (2000) and Hitchcock and Woodward (2003a, 2003b). For instance, whilst Woodward (2000) pays close attention to the identification of 'domains of invariance' in place of subsumption under law he all but ignores the piecemeal nature of model construction.

not only of why causation is important to science and how it is to be investigated but also of what use a workable causal theory can be put[19].

Even so, the differences should not overshadow altogether the resemblance Pearl (2000a) bears to the causal/mechanical project of explanation in science and the semantic conception of theories. There are some striking similarities between Pearl's formal approach to modelling, and the idea that scientific theories are abstract structures that are related to phenomena by some form of mapping relation. Moreover, the causal/mechanical tradition is a comparatively broad church. Those who consider van Fraassen's (1980, 1989) position on explanation in science to be consistent with the causal/mechanical project will see no harm in taking Pearl's account of modelling causes to be consistent also[20]. Furthermore, the semantic conception has continued to evolve alongside the causal/mechanical tradition; Suppe (2000) argues for continuity between contemporary research on the nature of scientific theories and his (1977, 1989) and Suppes (1962), with the exception that it is models that carry the burden rather than theories (Suppe 2000: 109-110). It is fair to say that debates about explanation in the philosophy of science are far more diverse today than compared to the time when Carnap and Hempel dominated the project of accounting for scientific explanation. An increasing number of philosophers take interest in debates that centre on modelling practices in science and its relationship with explanation[21]. Whether or not the divergence is a result of a difference of perspective on scientific methodology accompanied by a difference of attention regarding the question of what it is to model a given (causal) system would

---

[19] I suggest that this divergence between research in science and the philosophy of science is widespread. However, I do not thereby intend to suggest that Philosophers and Scientists aren't in agreement due to the fact that they belong to incommensurable fields of study. Indeed, naturalistically inclined philosophers tend to think that philosophy and science are continuous. Instead, I think that philosophers and scientists simply hold different aspects of problems to be important as a matter of contingent fact. For instance, see Pearl's statements identifying the 'quest for understanding' (philosophy's why questions) with how questions and hence with successful prediction under variable circumstances (Pearl 2000: 26).

[20] See also Suppe (1989).

[21] But where, again, possibly as a result of a difference in attention, philosophers often assimilate the role played by causation in causal modelling uncritically with an account of explanation tied to laws of nature. But see Woodward (2002a) and Hausman and Woodward (1999) for an account that does not do this.

take some time to illustrate[22]. In any case, that a gulf exists is clear, even if I have not succeeded in characterising its features correctly[23].

Recent research into both the nature of scientific models and how models are used to aid scientific reasoning is presented in, for example, Magnani and Nersessian (2002), Norton and Suppe (2001) Morrison and Morgan (1999), Magnani *et al.* (1999) and Herfel *et al.* (1995). Models are commonly employed in the sciences as analogies, as parts of cognitive systems and as representation devices (Giere 1999). Some argue that models, experiments, and simulations are not categorically different to each other since each can function as an object of investigation in its own right. Along these lines Boumans (2002) argues some models are not even descriptions or representations. Instead some models perform the role, not of a representational entity, but of a data sensor where the model is central in the creation of data. Displacing the primacy of isomorphism or similarity as the relation between model and reality sidesteps the need for bridging principles and ceteris paribus clauses opening a space for what Boumans and others call 'negligibility assumptions' in the application of models to problems (Boumans 2003: 317).

Glennan (1996, 2000a, 2000b, 2002) offers a philosophical account of models and mechanisms that has much in common with Pearl's conception of causal models. As such it is worth spending some time to articulate the salient portions of Glennan's (2000a) account of scientific models toward the goal of setting out an orienting context within which to interpret Pearl's account of causality. A caveat is appropriate here. Granted Glennan's account is instructive, I do not intend that it serve as a direct translation of Pearl's account and its commitments nor does Glennan offer it as one. Even so, there are good reasons for using Glennan's account. One is that Glennan's account of mechanism is the one philosophical account that is most consistent with Pearl's view of modelling practice together with Pearl's conception of structure.

---

[22] Hitchcock (2001: 641) briefly addresses this issue.

[23] For instance, philosophical research on the relation between theory and experiment in science does not sit altogether comfortably with the divergence. See Hacking (1991) for discussion. It is arguable

Glennan (2000a) aims to provide an appropriately general account of scientific models. He calls his account a 'model of scientific models' as distinct from a 'theory' of scientific models in the spirit of highlighting the practical aspect of modelling in science. At the centre of Glennan's (2000a) account stand his definition of a 'mechanical model' (hereafter (MM)) and the concomitant notion of a 'mechanism' (M):

> (MM) A mechanical model is a description of a mechanism that includes (i) a description of the mechanisms behaviour and (ii) a description of the mechanism that accounts for that behaviour.
>
> (M) A mechanism underlying a behaviour is a complex system which produces that behaviour by the interaction of a number of parts according to direct causal laws[24].

The details of the account are as follows. According to Glennan's (2000a) 'model of scientific models,' mechanisms are assumed to underlie 'behaviours'. The 'behaviour of the mechanism' is taken to be what the mechanism does. Direct causal laws are defined in Glennan's account as counterfactual-supporting generalisations that describe how changes, whether spontaneous or not, in one part of the model directly produce changes in another part. The 'directness' of the laws is a requirement that pertains to the efficiency with which mechanisms, and so models, are described. That is, the mechanisms that figure in a model should not be described at the expense of the model's completeness. There is, therefore, some sense in which 'direct causal laws' have an atomic description. The notion of law referred to in (M) picks out that class of generalisations made true by the nature of the contingent facts about how mechanisms are constituted and configured. The behaviour of a mechanism then

---

that a related divergence exists between the philosophy of logic and logic in Artificial Intelligence research. For discussion of the relationship between logic and AI see Thomason (2003).

[24] Glennan (2000) claims this conception of mechanisms can be found in the work of Winsatt (1974) and Herbert Simon.

supervenes on (a class of) counterfactual supporting generalisations. In particular, Glennan expects that the reliable behaviour of mechanisms depends upon the existence of lawful relations between their parts, and direct causal laws characterise these relations[25]. It should be noted that Glennan's notion of laws is significantly different from the notions of laws that figure elsewhere in philosophical conceptions of scientific theories. For instance, Glennan's account is to be contrasted with the conception of laws that figures prominently in the 'covering law' account of explanation. For Glennan, laws, as these pertain to scientific models, do not cover most ordinary phenomena as laws of nature are expected to. For instance, laws on Glennan's account have narrow scope, whereas the standard account takes laws to have wide scope. Furthermore, there are innumerable laws, as many as one for each mechanism on Glennan's account whereas, according to the covering law conception of laws, there are expected to be few *bona fide* laws of nature (Glennan 2000a: 10-11)[26].

Note that Glennan's characterisation of mechanical models has two parts–one concerned with the description of the behaviour of the mechanism, and the other with its mechanical description. According to Glennan, the latter description pertains to the internal structure of the mechanism and the former to the external description. Descriptions are considered to be semantic entities akin to propositions. Hence, there may be many different formulations of the one descriptive entity and yet no change in the mechanism described, but a change in the specification of the descriptive entity necessarily involves the description of an alternative mechanism. On Glennan's model of scientific models the relation that obtains between a model and a mechanism is one of 'approximate similarity'. As Glennan remarks:

---

[25] There are many accounts of mechanism available in the philosophical literature. For instance, besides Glennan's account, see Salmon (1984), Machamer (2002), Machamer et al. (2000) and Woodward (2002a).

[26] Pearl (2000a) takes a similar notion to Glennan and adds further constraints. In particular, Pearl (2000a) considers only non-backtracking counterfactuals and ties these in turn to the notion of invariance. I discuss this greater length in section 2.0.3 and again in chapter 3.

> The behaviour of the system in nature is described (to varying degrees of approximation) by the model's behavioural description and the internal structure of the system is described (again to varying degrees of approximation) by the model's mechanical description. To make claims about the nature of a mechanism, one constructs a model and asserts that it is similar to a system in nature (Glennan 2000a: 12).

Glennan's account makes two further presuppositions relevant to the explication of Pearl's account. First, the concept of a mechanism's behaviour presupposes a concept of 'normal functioning'. The idea here is that, in describing the behaviour of a mechanism, the description is carried out on the supposition that the mechanism is not broken and this includes mechanisms that are not the product of design or selection. The second presupposition involves the relationship between behavioural and mechanical descriptions. It is assumed that this relationship is one-many due to the fact that the same behaviour can be produced by numerous distinct mechanisms. That is, behaviours underdetermine mechanisms. To Glennan this one-many relationship raises the following question: 'If one has two competing models of a mechanism which both predict the same behaviour, how does one choose?' (Glennan 2000a: 14). Pearl deals with this question via the specification of a 'minimality' constraint on the construction of causal models together with a 'stability' condition on model parameters and allows the investigator recourse to causal information not explicitly displayed by a given model. Moreover, for Pearl (2000a) the behaviour of a mechanism is described via a probability distribution and the mechanical description is carried by the functional equations and accompanying pictorial graph of a mathematical object labelled a 'structural causal model'. For now the details of these conditions and objects are not important; rather it is important to recognise that the overall account of models offered by Glennan (2000a) sets in place many of the key ideas according to which an interpretation of Pearl's account of causality may proceed[27].

[27] Glennan continues to develop his account of both mechanism and models. See for instance, Glennan (2000b, 2002) and Tabery (2004). I discuss these issues in greater detail in section 3.4 of chapter 3.

In any case, I began with the claim that the overall aim of the present research effort is to set out and to examine Pearl's theory of causation. I conclude this section with the thought that Pearl (2000a) does not intend to offer a theory of causation so much as a theory of causal modelling.

I have reasoned that the divergence is important to recognise since it affects what one thinks Pearl's theory is and, therefore, how criticisms of the theory are to be framed and a coherent interpretation traced. There is a considerable amount of detail that needs to be added to fully illustrate the nature and extent of this divergence. However, I have said enough to identify it and since doing that much has served present purposes I leave further discussion of its details to later chapters.

1.2 The Two Questions of Causation

In this section I take up the task of providing some detail to the enterprise of causal modelling in science in as much as the enterprise is conceived of by Pearl (2000a).

Pearl sets out his picture of causal modelling as follows[28]. The causal modeller may assume

> [...] that the world is described by random variables, some of which may have a causal influence on others. This influence is modelled by a set of structural equations, where each equation represents a distinct mechanism (or law) in the world, one that may be modified (by external actions) without altering the others. In practice, it seems useful to split the random variables into two sets, the exogenous variables, whose values are determined by factors outside the model, and the endogenous variables. It is these endogenous variables whose values are described by the structural equations (Halpern and Pearl 2001a).

A number of points from this comment require clarification. Pearl views the task of causal modelling as an induction game that scientists play against Nature (Pearl 2000: 43-45). The claim here is that Nature is assumed to act as though it were in possession of what Pearl calls 'stable causal mechanisms', which, were they open to inspection, would best be described as deterministic functional relations between (sometimes unobservable) variables. Furthermore, the modeller assumes that these mechanisms are organised in the form of an acyclic structure and then attempts to identify these mechanisms and/or specify the values of endogenous variables from available observations[29]. But all that is available to the modeller are data. What a

[28] Pearl (2000a) indicates that the discussion of causal modelling takes an idealized view both of model construction practices and in some instances the statistical methodologies. However, considerable detail is added as the account unfolds.

[29] This restatement is an idealization of the task of causal modelling. There are alternative modelling procedures that do not assume acyclicity and where assumptions regarding Nature's features differ significantly. I return to this issue when I discuss model identifiability and justification in section 3.4 of chapter 3.

causal modeller attempts to do is identify causal relations using non-experimental techniques. The point can be better illustrated after discussing the nature of experimental and non-experimental techniques.

In the context of the collection of sciences listed above one can distinguish two basic types of experiment; controlled and randomised. It is commonly accepted that each of these types is designed to aid in establishing the existence of cause-effect relationships. Since most view the randomised experiment as the stronger of the two (despite the fact that, as Shipley points out, the controlled experiment precedes the randomised experiment in terms of historical discovery) I will discuss it first (Shipley 2000: 7). To illustrate the logic of this experimental form, consider the following example due to Shipley (2000):

> [Picture an] experiment designed to determine whether the addition of a nitrogen-based fertiliser can cause an increase in the seed yield of a particular variety of wheat. A field is divided into 30 plots of soil ($50cm \times 50cm$) and the seed is sown. The treatment variable consists of the fertilizer, which is applied at either 0 or 20kg/hectare. For each plot we place a small piece of paper in a hat. One half of the pieces of paper have a '0' and the other half have a '20' written on them. After thoroughly mixing the pieces of paper, we randomly draw one for each plot to determine the treatment level that each plot is to receive. After applying the appropriate level of fertilizer independently to each plot, we make no further manipulations until harvest day, at which time we weigh the seed that is harvested from each plot.
> The seed weight per plot is normally distributed within each treatment group. Those plots receiving no fertiliser produce 55g of seed with a standard error of 6. Those plots receiving 20kg/hectare of fertilizer produce 80g of seed with a standard error of 6 (Shipley 2000: 7).

The randomisation of the treatment allocation allows the researcher to calculate the probability that the results occurred by chance and, given a considerably small

probability in this instance, distinguish between chance associations and systematic ones[30]. Systematic associations, in contrast with chance associations, are assumed to be the result of some underlying mechanism (Shipley 2000: 7). When the probability that a chance event has occurred is sufficiently small the researcher discards the possibility that a rare event has occurred and, in the present example for instance, reasons that there is very good evidence of a positive association between the addition of fertilizer and the increased yield of the wheat.

Many statisticians and scientists accept that the process of randomisation allows them to differentiate between associations due to causal effects of the treatment and associations due to some variable that is a common cause of both the treatment and response variables. Hence, the next step in the process is to examine what reasons there may be for concluding that the result is in fact due to causality. It is generally accepted that there can only be three basic causal explanations of an association between two variables X and Y: either X is the cause of Y, Y is the cause of X, or there are some other causes that are common to both X and Y. In the example above, Shipley asserts that we can exclude the possibility that seed produced by the wheat caused the amount of fertilizer that was added since it is clear in this instance that causes must precede their effects in time. Moreover, we have to hand the common causes of treatment quantities and treated plots. These are simply the numbers that the experimenter saw written on the piece of paper attributed to an actual plot. However, it remains a possibility that these are not the sole common causes in operation. But, even so, Shipley continues, we can exclude the possibility that the observed association is due to unrecognised common causes since, by definition, the random process by which the treatment units were chosen ensures that the order in which the plots receive the treatment is causally independent of any attributes of the plot, its soil, or the plant at the moment of randomisation (Shipley 2000: 8).

Shipley summarises the logic of the randomised experiment in the following way.

---

[30] Shipley (2000) calculates that the probability that a rare event occurred is 5×10^{-8}.

> We began by asserting that, if there was a causal relationship between fertilizer addition and seed yield, then there would also be a systematic relationship between these two variables in our data: causation implies correlation. When we observe a systematic relationship that can't reasonably be attributed to sampling fluctuations, we conclude that there was some causal mechanism responsible for this association. Correlation does not necessarily imply a causal relationship from the fertilizer addition to the seed yield, but it does imply *some* causal relationship that is responsible for this association. There are only three such elementary causal relationships and the process of randomisation has excluded two of them. We are left with the overwhelming likelihood that the fertilizer addition caused the increase in seed yield (Shipley 2000: 9)[31].

Hence, the process of randomisation serves two purposes in causal inference. First, randomisation causally isolates the experimental units from the treatment variable and other possible common causes. Second, randomisation helps the investigator to reason about the likelihood that an association is due to chance and not to a causal mechanism[32] (Shipley 2000: 10; Pearl 2000a: 347-348).

The other basic form of experiment is the controlled experiment. A controlled experiment consists of testing hypothesised causal relationships by deducing what would happen if specific variables in the experiment were fixed in a particular state (controlled) and comparing the result of the deduction with the observed result (Shipley 2000: 15). Holding a variable fixed has generally meant taking some steps to

---

[31] The alternative causal explanations are not excluded categorically since it is possible that in this case the plots that received the fertilizer had some attribute, such as higher moisture holding capacity, that actually caused the increase in seed yield. See Shipley (2000: 9-10) for discussion in the context of this example.

[32] It is usual that randomised and controlled experiments have a population significantly greater than 1. But this is not always the case. In some cases of medical diagnosis, for instance, an experimental method, called Single Patient Outcome Trial (SPOT), is used to ascertain the effects of interventions on just one individual over time. The aim of the method is to discover and catalogue how such interventions affect the individual. The method is contrasted with diagnostic techniques and treatments that extrapolate from patterns described by studies carried out on populations. I note that the method resembles Pearl's view of actual causality.

physically intervene in the experimental set-up so that they can no longer vary naturally. However, control in an experiment is not necessarily physical and may often be observational (statistical) (Shipley 2000: 15-16). The latter form of control involves the notion of statistical conditioning and leads naturally into a discussion of non-experimental methods of identifying causal relationships. The basic conception of causal inference within the non-experimental domain involves the construction of a language within which to express causal claims and a translation procedure for moving between the language of causality and the language of statistics (i.e. probability). The idea is that in cases where manipulation, control and randomisation cannot actually be implemented the queries held by an investigator might nonetheless be translated into a model, which then serves as a simulation of the experiment the investigator intended to perform. In this sense, the investigator attempts to set out the necessary and sufficient conditions needed to specify a joint probability distribution that must exist given a specific causal process (Shipley 2000: 24-25). In taking the step away from experimental methods, the investigator makes causal claims on behalf of the mathematical details of the model in conjunction with the investigator's knowledge, rather than on behalf of the experimental design. Pearl likens this step in the modeller's task to an attempt to acquire causal knowledge without the benefit of randomisation or control (Pearl 2000a: 42-43). What allows the modeller to make the inference to causality in these instances involves the identification of what Pearl calls *autonomy*. I discuss the details of autonomy below. What is important to note at this point is that Pearl denies modellers must give up on causality just because they cannot conduct experiments. The reason for the denial is that Pearl thinks of randomisation and control as mere markers of autonomy (Pearl 2000a: 63, 253). Hence, it is not experimental design in and of itself that licenses causal inference. It is the fact that randomisation and control are ways of tracking autonomous mechanisms that endows them with causal characteristics.

From this conception of the causal modeller's main occupation it is clear why Pearl thinks that the fundamental questions of causality are:

(1) What empirical evidence is required for legitimate inference of cause-effect relationships?
(2) Given that we are willing to accept causal information about a phenomenon, what inferences can we draw from such information and how can we draw them?

In effect, Pearl is claiming that causal relationships are common in nature but often complex and obscure. That there is causation in nature is taken as granted[33]. The (difficult) task that remains is its identification. By analogy, Pearl's view of causal modelling is akin to the attempts an investigator might make to construct a machine that behaves (over time or under interventions) in the same or relevantly similar ways to the part of reality the investigator has under investigation. There are several assumptions Pearl claims we are justified in making about the world the investigator seeks to understand. The investigator is seeking to learn cause-effect relations via uncontrolled observations of Nature. Pearl insists that humans demonstrate the ability to learn cause-effect relations day-to-day and so he intends to craft tools, which in providing a means to answer the following three questions, allow for the codification of this ability:

1. What clues prompt people to perceive causal relationships in uncontrolled observations?
2. Is it feasible to infer causal models from these observations?
3. Would the models inferred tell us anything useful about the causal mechanisms that underlie the observations?

Pearl's picture of the investigator's task first assumes that the world as seen through the investigator's eyes is (quasi-) deterministic. In Pearl's words, his view of causal models is that they be understood to:

[33] In as much as the causal modeller is concerned. However, Pearl has a position regarding the epistemology and perception of causal relations, which I discuss in section 3.3 of chapter 3.

> ... reflect Laplace's (1814) conception of natural phenomena, according to which nature's laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions (Pearl 2000a: 26).

Pearl bases his preference for a Laplacian conception of (the modelling of) Nature on three considerations. First, the Laplacian conception of Nature is more general than the competing stochastic conceptions[34]. Second, the Laplacian conception of Nature is closer than the stochastic conceptions to human intuitions[35]. Third, counterfactual propositions are easily defined within the Laplacian conception but are not definable within a stochastic conception due principally to the fact that the probability calculus (alone) cannot express the structure of counterfactual expressions (Pearl 2000a: 26-27).

I do not think Pearl intends this to be a proclamation concerning the actual nature of the physical world although it does belong to a specific world-view. Instead the claim of determinism is linked to the role played by the Markov condition in causal modelling and, as such, is thought of as a convention guiding the causal modelling procedure. As is mentioned in the introduction the Markov condition (for Pearl) states that the value of a variable in a causal model is independent of that variable's effects, conditional on its direct causes. To see this, consider the following overview of the modelling procedure. In one part of the modelling procedure the causal modeller takes the following steps. First, the modeller builds a prototype model structure designed to meaningfully embody causal claims, which Pearl calls a 'causal structure'. This structure consists only of a set of variables such that each distinct variable is represented as a node within the graph, and where the edges of the graph represent direct functional relationships among corresponding variables (Pearl 2000a: 44). Second, the modeller then attempts to construct from this prototype a workable

[34] I take Pearl to mean by 'stochastic conception' the view that Nature is inherently probabilistic and that without the benefits that accompany knowledge of 'laws' or 'regularities' the modeller must accept that each event is equi-probable.

[35] This assertion is underpinned by Pearl's belief that humans store causal information in terms of counterfactuals and his view that such causal information is stable under the influence of external interventions.

causal model by specifying the precise way in which each variable is influenced by its parents in the graph.

In Pearl's words, when building a model the

> [...] Markov condition guides [the investigator] in deciding when a set of parents is considered complete in the sense that it includes all the relevant immediate causes of variable X. It permits [the investigator] to leave some of these causes out of the set of parent variables (and be summarised by probabilities), but not if they also affect other variables modelled in the system (Pearl 2000a: 44)[36].

Most importantly, the use of the Markov condition reflects the *assumption* that Nature has a structure such that correlations are sometimes reflective of a deterministic causal structure[37].

However, Pearl further assumes that:

> ... Nature is at liberty to impose arbitrary functional relationships between each effect and its cause and then to perturb these relationships by introducing arbitrary (yet mutually independent) disturbances. These disturbances reflect 'hidden' or unmeasurable conditions and exceptions that Nature chooses to govern by some undisclosed probability function (Pearl 2000a: 44).

The upshot is that the causal modeller can access only a subset of variables - the ones that are measurable or observable - since, it is thought, Nature hides the full detail of

---

[36] I discuss the Markov condition and its status within Pearl's account at greater length in chapter 3.
[37] I note here that Pearl does not explicitly draw any distinction between determinism and determinability. However, the distinction is treated by fiat since Pearl considers making predictions under conditions of uncertainty of a type related to lack of information. But, on several occasions Pearl conflates (epistemic) predictability with (ontic) determinism. I take this issue up in chapters 2 and 3 where I find that the primary sense given to determinism is the one in which certain variables of a structural equation are a function of other distinct variables of the same equation.

the underlying causal structure of the phenomena being modelled. It seems to follow that with only observed or measured variables to hand the modeller can only construct a model that defines a joint probability distribution over the variables in a system. But since the observed variables are but a subset of the variables of the system (the remainder being unmeasured) the resultant model may simply be on a par with a (possibly infinite) class of observationally equivalent models.

So far then the 'machine' constructed by the investigator may have a 'good fit' with the phenomenon. But this is only in the sense that the machine is able to (more-or-less) match the behaviour of the set of variables open to measurement to the satisfaction of the person who constructed the machine. The idea of getting the machine to match the behaviour of the variables conjures notions of simulation. Two metaphors for simulation cited include building the machine to 'look like' the data on the one hand and building the machine to 'operate' like the process that generated the data. The latter metaphor is the more apt of the two for causal modelling[38]. Either way, the metaphor of the machine and the discussion of causal modelling that precede it serve to set the stage for the interpretation of Pearl's account of causality that follows.

[38] See, for instance, chapter 1 of Shipley (2000).

1.3 Manipulation, Intervention and Agency

In this section I examine how Pearl's (2000a) account has been interpreted in recent literature. At the time of writing the number of articles and books containing specific comment about Pearl (2000a) are relatively few. The literature does, however, span several disciplines and continues to grow as Pearl's ideas are more widely disseminated and as researchers attempt to put his ideas to use. In the philosophical literature on causation it appears an early orthodoxy has formed around the opinion that Pearl (2000a) offers what is essentially a 'manipulation' account of causation and has renewed focus and attention on the so-called 'agency' theories of causation. After introducing comments from economic and epidemiological literature I spell out the characteristics of agency theories of causation and briefly discuss the strength of the view that Pearl is a manipulation theorist.

Providing a detailed interpretation of Pearl's theory is no mean feat. The task is not aided by the fact that Pearl (2000a) never actually specifies a complete stand-alone account of his theory despite the overall thrust of the publication being towards its elucidation. Pearl expends considerable effort elaborating on how his account is best interpreted – the fundamentals of which I have attempted to present in sections 1.1 and 1.2. But, despite Pearl's best efforts enough leeway exists in his presentation for several distinct interpretations to exist. A survey of published comment regarding Pearl (2000a) shows his theory to be variously interpreted. Differences in opinion concerning the key components and aims of the theory typically vary according to the central questions and debates of the disciplines each reviewer represents. For instance, econometricians have taken the main thrust of Pearl (2000a) to concern methodological issues. LeRoy (2002), for example, asserts:

> Three ideas underlie everything Pearl writes:
>
> > (1) Causal ideas are indispensable in [the natural and special sciences, philosophy, and statistics]. Pearl rejects the calls one periodically

> hears to dispense entirely with the terms 'cause' and 'effect' or substitute for them terms like 'functional dependence', as if by doing so one could somehow circumvent the need to deal explicitly with causal ideas.
>
> (2) Causality is different from probability. Even though the two are obviously related, they are not identical, and a separate analysis is required if probabilistic relations are to be interpreted causally.
>
> (3) The informality with which most of us use causal language leads to much confusion, and this confusion could be avoided if we made more use of formal methods to analyse causality. Indeed one purpose of [Pearl (2000a)] is to convince us that the relevant formal tools - principally graph theory - are already available and well developed.
>
> For what it is worth, I completely agree with the first two points and the first part of the third. As to whether graphical analysis, or formal methods generally, have much to contribute to the analysis of causation as Pearl believes, I am not yet convinced, particularly with regard to economics[39].(LeRoy (2002)

Some epidemiologists comment that graphical modelling methods such as Pearl's

> [...] have seen extensive analytic application (especially in the social sciences) [but] nonetheless, in epidemiology these models remain confined largely to the conceptual teaching realm (to the extent that they appear at all) (Greenland and Brumback 2002: 1036).

[39] If graphical methods are indeed the fundamental component of Pearl's theory, then LeRoy's comments bear some weight, since not being convinced of the formalisation is, in effect, not to be convinced of Pearl's theory of causality. What remains on the table for LeRoy then are just methodological concerns relating to study design.

Other epidemiologists comment that the importance of Pearl (2000a) lies in its unification of various different approaches to causal modelling. According to these epidemiologists Pearl (2000a) has shown how one or another approach to expressing and representing causal relationships (such as Structural Equation Models or causal diagrams) translates directly into other approaches (such as counterfactual models) (Maldonado and Greenland 2002).

Philosophers tend to agree amongst themselves[40] that Pearl (2000a) elaborates a manipulation account of causation. Whether or not this opinion is well justified it is in any case clear that philosophers have taken Pearl to offer an *analysis* of causation. For instance, commenting on a quotation from Cook and Campbell (1979) Woodward (2001) asserts:

> [Cook and Campbell's general ideas regarding causation as manipulation] are commonplace in econometrics and in the so-called structural equations or causal modelling literature, and very recently have been forcefully reiterated by the computer scientist Judea Pearl in an impressive book length treatment of causality (Pearl 2000a) (Woodward 2001: 1).

Such that, Woodward thinks, "[t]he characterisation of the notion of intervention is rightly seen by many writers as central to the development of a plausible version of a manipulability theory [and] [o]ne of the most detailed attempts to think systematically about interventions and their significance for understanding causation is due to Pearl (2000a) [...]" (Woodward 2001).

But, for others, Pearl is not primarily engaged in developing a manipulability theory of causation so much as offering a counterfactual theory. For example, Menzies (2002) asserts that Pearl (2000a) articulates a counterfactual theory of token causation:

---

[40] But this is by no means universal. For instance, see Hitchcock's review in Hitchcock (2001). See also Menzies (2002).

> Pearl's theory of token-causation can be called a counterfactual theory. In his (2000), he attempts to capture within the structural equations framework the notion of quasi-dependence that Lewis (1986) introduced as a tentative solution–though later discarded–to the difficulties that the late pre-emption examples posed his original counterfactual theory (Menzies 2002: 4).

Variation of opinion concerning the nature of Pearl's project is not surprising given the differing histories and developmental contingencies within and between disciplines. But, this is not to say that there is no (cross-discipline) agreement about the nature of Pearl's project. As I mentioned above, it is still early days for the analysis of Pearl's account. Two points are worth making. First, it cannot be taken as given that Pearl has articulated a stand-alone theory of causality. Second, differences in the various disciplinary comments on Pearl (2000a) together with the fact that Pearl has attempted to write for a somewhat heterogenous audience have a strong bearing on what the key components of the theory are taken to be. Broadly speaking, most agree that Pearl attempts to account for causality by using the properties of path diagrams of graphs. But, even so, the details of Pearl's theory depend upon whether one endorses their applicability. In this sense opinion varies according not just with the details of Pearl's account but also with what domain of problems Pearl intends his account to be useful in solving. Across the literature on causality these tasks vary from aiding investigators in drawing causal inferences, to the identification of confounding, to the metaphysical analysis of causal relations. The breadth of this domain of application is one explanation for the variety of different opinions concerning the nature of Pearl's account. Writing for a heterogenous audience has this disadvantage.

Another aspect of this plurality involves the level at which Pearl's account is pitched. Pearl (2000a) throws into question several foundational issues not typically addressed by statisticians, epidemiologists, and econometricians. For instance, Pearl asks econometricians and statisticians why they resist formal approaches to assessing

causal claims. Likewise, Pearl notes that confounding bias is somewhat of an anathema to statisticians because accounting for it requires reckoning with causality. Statisticians might argue that in so doing Pearl steps into philosophical or 'metaphysical' territory. They may be correct. Pearl's views on foundational issues, sparing as they are, extend to (the discipline of) philosophy, where Pearl has questioned what the discipline is meant to contribute[41].

I now turn to examine more closely the contribution made by philosophers towards an interpretation of Pearl's account. As I mentioned above the leading interpretation categorises Pearl as a manipulation theorist. The details of manipulation theories and the claim that Pearl is a form of manipulation theorist are as follows.

Psillos (2002: 6-8) recently elaborated on a distinction between what he calls the platitudes of causal theorising and our pre-theoretical intuitions about the nature of causation. The four platitudes he cites include the difference platitude, the recipe platitude, the explanation platitude and the evidence platitude[42]. The pre-theoretical intuitions tend toward one of two centres; the apparent intrinsic nature of the causal relation on the one hand and the intuition that causation must exhibit regularity on the other. I wish to draw particular attention to the recipe platitude and use it to introduce the notion of manipulation and its place in causal theorising. As for the distinction itself, Psillos's aim in introducing it is to highlight the disparities that exist within the field of causal theories and to utilise it as a means to break up the field of causal theories into instructive categories[43]. At the most general level Psillos's categorisation of causal theories distinguishes between those that entertain the existence of causal powers or dispositions of some form or other and those theories that do not. Speaking broadly, those of the former category are labelled non-Humean theories of causality and those of the latter category Humean or regularity theories of causality. The

[41]On the issue of what philosophy contributes see Pearl's challenge in section 3.4 of chapter 3. See also Pearl (2000a: 310).
[42] See also Mellor's discussion of 'connotations' in his (1988: 230) and discussion in chapter 1 of Dowe (2000).
[43] For an alternative view on categorising theories of causation see Menzies (1999).

several platitudes are then mobilised either individually or in combinations to act as yardsticks against which a theory from one category or the other may be compared.

According to the recipe platitude, causes are a means to producing or preventing their effects[44]. In more modern guise the recipe platitude is a statement of the fact that causes can be used to manipulate their effects but, typically, not vice versa. This idea has a relatively rich history and there is some considerable detail afforded to it in the literature.

For instance, the later Wittgenstein arguably came to the view that causation was a 'family resemblance concept' in as much as it proved to be immune to specific analysis. That is, causation for Wittgenstein reflects an interconnected web of concepts related by similarity, which may be taken account of via mapping, rather than by an analysis conducted in terms of, say, necessary and sufficient conditions. Focusing on how people establish causal connections in day-to-day contexts, Wittgenstein noted a variety of 'prototype' causal connections. These include impact; traction; mechanism; human reactions to sensation and emotion; and regularity of succession (Glock 1996: 72-73). Such prototypes are, according to Wittgenstein, manifest in the variety of human practices. Hence, no one prototype in particular was thought by Wittgenstein to be fundamental. Even so, it seems clear Wittgenstein thought that the notion of causation as agency was at least genetically prior to that based on observation[45].

Gasking took causation to be essentially related to the manipulative techniques humans employ in order to produce a result. Influenced to some degree by Wittgenstein's thought on the subject, G.H. von Wright claimed that

> ...to think of a relation between events as causal is to think of it under the aspect of (possible) action. It is therefore true, but at the same time a little

[44] The term 'recipe' in relation to causes is due to Gasking (1955).
[45] For further discussion see Glock (1996) and Wittgenstein (1976).

> misleading to say that if p is a (sufficient) cause of q, then if I could produce p I could bring about q. For that p is the cause of q, I have endeavoured to say here, means that I could bring about q, if I could do (so that) p (von Wright 1971: 74).

That is, von Wright held that nature would unfold in a deterministic fashion but for the interference of agents. It is by the action of agents that an analysis of causation becomes a possibility. Hence, for von Wright "p is a cause relative to q, and q an effect relative to p, iff by doing p we could bring about q, or by suppressing p we could remove q or prevent it from happening" (Sosa and Tooley 1993: 16). Moreover, the connection between causation and action is logical in its nature and as such is intertwined with the raft of human practices that surround free action (von Wright 1973).

In more recent work, Menzies and Price argue that causation should be analysed as a secondary quality. Their basic idea is that "... an event A is a cause of a distinct event B just in case bringing about the occurrence of A would be an effective means by which a free agent could bring about the occurrence of B" (Menzies and Price 1993: 187). Menzies and Price (1993) in effect are saying that causation is reducible to the volition of agents to 'bring about' some effect via the manipulation of its cause, and that the notion of an agent 'bringing about' an effect is based upon non-causal (non-linguistic) ostension (similar to the role that ostension takes in the analysis of colour concepts) (Psillos 2002: 103).

Woodward's account of (causal) explanation in scientific contexts emphasises the notion of intervention. For Woodward the relationship between two events (variables), call them X and Y, can be considered causal if, based on an intervention that changed the value of X, the value of Y would change whilst the overall relationship between X and Y remains stable (Woodward 2000: 205). This formulation involves a counterfactual component. What this means is that the notion of intervention is not limited to interventions that are actually performed. Instead the

idea is that X is a cause of Y if and only if Y would change if an appropriate manipulation of X were carried out. According to Woodward (and others), Pearl offers a similar type of interventionist account of causation. This claim is primarily based on the fact that Pearl (2000a) utilises causal graphs that, when constructed according to the principles and assumptions Pearl recommends[46], act as 'oracles'–a counterfactual predicting device–that display the effects of interventions.

Each of these views draws attention to the fact that causation[47] essentially involves the notion of manipulation. As such each view places the recipe platitude somewhat ahead of the others in order of importance. Even so, each view differs significantly from the others. Manipulation theorists have, often to avoid inheriting the shortcomings of a predecessor or in response to criticism, introduced several distinctions that help to differentiate their specific view from others of the same variety. Some of these are worth outlining since a number will be instructive when the time comes to detail and assess Pearl's account of causality.

By adding another level of detail on top of the general manipulation thesis, one can claim that causation involves manipulation conceived in terms of either free human action or independently of human action[48]. Analyses of causation pertaining to the former view have come to be called 'agent' accounts and those pertaining to the latter are called 'interventionist' accounts. Clearly, Gasking, Menzies and Price (1993), and less clearly von Wright stand under the umbrella of the agency account since it is part of their views that causation requires essential reference to (human) agency. But, Woodward (and Hausman) stand with the interventionist's since they seek a definition of causation that is not reducible to an agent's volition or intention.

Still further divisions exist. Manipulation accounts of causation may seek to be reductive or not and may attempt or decline any attempt to offer a universal account

[46] Primarily the notions of stability, invariance, and dependence, which I detail in chapter 3.
[47] And in some instances explanation also.

of the causal relation. In the case of the former, a reductive manipulationist account of causality asserts that causal relations can be completely explained in terms of manipulation or intervention. In the case of the latter, the manipulation theorist asserts that there are no causal relations that a manipulation theory cannot take account of. Further differences amongst individual manipulationist views of causality may follow according to the stance they adopt toward the remaining connotations of causation; the order with which one thinks answers should be provided to questions concerning causal inference as distinct from causal analysis; how the account relates to laws of nature; and what the account makes of instances of singular or token causation.

However, to my mind there remains some question as to whether Pearl should be classed as a manipulation theorist since, in light of Pearl's intentions, it is not clear that he means to articulate such an account, and even if it turns out that he has (possibly unintentionally) done so, it is not true that manipulation is the only available category to which his theory might belong nor is it clear that manipulation is at the centre of the theory[49]. The primary reason uncovered so far for denying Pearl the title of manipulation or agency theorist is that despite Pearl's obvious endorsement of the means/end intuition, Pearl holds that manipulations only reveal causal relations when they are in fact there to be revealed. Hence, it seems clear that Pearl would not accept that 'causality means manipulation', much less that causation *is* manipulation. Also, it is not clear that one can move from conceptions of manipulation in experimental settings to manipulation of model structures without loss or change of meaning. This will not, however, be the last word on the matter. I return to this question in chapter 2.

On the issue of differing standpoints I declare my interest in Pearl's account to lie within the discipline of philosophy. I take it that Pearl (2000a) contains an analysis of

[48] Taking this platitude seriously might lead one to manipulation but other commitments might also lead one into agreement with the recipe latitude. It can be a starting point in one's causal theorising or a happenstance finishing point.

causality granted the caveats of sections 1.1 and 1.2 concerning causal modelling. In the final chapter I argue that Pearl offers a 'two-part' theory of causation that is at base neither manipulationist nor counterfactual. But, before I provide the details of my interpretation it is first necessary to set out the key components of Pearl's account as they appear in their original form.

[49] That is, in spite of comments to the contrary on p 104 regarding Spirtes's 1991 lecture at the International Congress of Philosophy of Science (Pearl 2000a: 104).

2.0 Pearl's Theory of Causality

There are several ways to trace out an account of Pearl's theory of causality. In fact there are several accounts available in the literature. The purpose of this chapter is to present and comment on Pearl's account of causality more or less as one finds it in Pearl (2000a). The presentation primarily draws on Pearl (2000a) and secondarily upon Halpern and Pearl (2001a)[50]. I intend the presentation to represent what the reader may view as a standard account of Pearl's theory. This account will then function as the reference point of the philosophical interpretation I offer in chapter 3. I deal with specific exegetical issues of importance in the later sections of this chapter and again in the following chapter as the need arises. One such issue is the need to clarify the role that counterfactuals play in Pearl's account. I do this in section 2.1.

Pearl contends that the path traversed through each chapter of his (2000a) roughly parallels that in which the account developed over time. What I find useful in tracing this path is the transition from Bayesian Networks to Causal Functional Models. However, several issues dictate that a clearer presentation might be had by detailing some of the key concepts and principles of Pearl's account out of step with their historical discovery. Hence, I will begin with Bayesian Networks and move to Causal Functional Models more or less in step with the progression in Pearl (2000a). But I will break step to present definitions and conditions that pertain to what Pearl calls 'actual causation'; Pearl's account of type and token causes; and Pearl's account of counterfactual statements. At a conceptual level the presentation traces a path that begins with the mathematical theory of graphs, especially Directed Acyclic Graphs, moves through the detail of Pearl's logic of causal reasoning and ends with Pearl's demonstration of how expressions of the logic may be translated between different forms of causal models in accord with the type of problem the modeller is attempting to solve. The presentation is broken into three components reminiscent of many

presentations of formal logical systems. First I set out the syntax and then I trace its semantics. Last, I present the informal reading of the semantics.

2.0.1 Syntax

It is fair to say that the foundation of Pearl's account is the Bayesian Network. A Bayesian Network is a form of path diagram or graph capable of representing joint probability functions over distributions of variables[51]. The so-called graphoid axioms confirmed by Dawid (1979), Spohn (1980), and Pearl and Paz (1987) determine the properties of Bayesian Networks[52]. As such, a Bayesian Network can be thought of as a mathematical object designed to represent and to facilitate reasoning about probabilistic relationships between sets of variables. According to Pearl, Bayesian Networks are so-called for three reasons. First, their input information is subjective, representing (in part) the investigators' knowledge. Second, Bayes's conditioning is used to update information in the light of new evidence; Third, decision procedures for Bayesian Networks distinguish between causal and evidential modes of reasoning[53] (Pearl 2000a:14). Bayesian Networks have several alternative applications to those which Pearl has put them. These uses include language understanding, map learning, and decision making. Bayesian Networks are known by several other names such as knowledge maps, (Bayesian) belief networks, expert systems and probabilistic causal networks (Charniak 1991: 50). Recently the use of Bayesian Networks has tended to be greatest in two fields of study: Artificial Intelligence research and the study of human decision-making (Kennett *et al.* 2001). Despite the variety of applications the following presentation is intended to track the

---

[50] I also draw upon a number of technical papers authored by Pearl and others. See citations in text.
[51] Or conditional probability tables.
[52] See Pearl (1987) and Geiger *et al.* (1990). See also Edwards, D. (2000) for a recent introductory treatment of graph theory. See Charniak, E. (1991) for an introduction to Bayes Nets. The graphoid axioms are recounted in Pearl (2000a: 11). I do not set out the graphoid axioms since they are not necessary for the discussion. The main thrust of the axioms and the notion of information relevance will be covered when introducing the d-separation criterion and independence assumption below.
[53] None of these conditions are necessary to define the properties of graphs.

use and conception of Bayesian Networks consistent with Pearl (2000a), which is primarily geared towards statistical and causal modelling.

Pearl asserts that there are two primary reasons for using graphical methods in probabilistic and statistical modelling. One is that Bayesian Networks are capable of perspicuously representing the investigators' assumptions (typically concerning structure) and the other is that Bayesian Networks offer an effective method of representing joint probability functions. In other words, Bayesian Networks "[...] provide a compact and clear representation of complicated probabilistic independencies" (Twardy and Korb 2002: 2) and offer a visual as well as formal guide to facilitate the drawing of inferences (Pearl 2000a: 13). The syntax of Pearl's account has two main components. The first is the language of graphs and the second is the so-called 'calculus of interventions', which in part is the rules of inference governing permissible alterations to graphs. To gain a reasonable picture of graphs, and so too Bayesian Networks, I will briefly set out the basic principles of graph theory and follow by commenting on the relation between statistical models and Directed Acyclic Graphs.

A graph is a mathematical object constructed from a set V of vertices (or nodes) and a set E of edges (or links) connecting some pairs of vertices. For Pearl, nodes correspond to variables and edges denote a relationship that holds between pairs of variables, the interpretation of which varies according to application. The edges in a graph can be directed or undirected. Directed edges are marked by a single arrowhead oriented toward the direction of influence and undirected edges remain unmarked. In some applications graphs may use 'bi-directed' arcs to denote the existence of unobserved common causes or confounders. If all edges in a graph G are directed the graph is called a 'directed' graph. Removing all arrowheads from the edges in a graph G results in an undirected graph called the 'skeleton' of G. A 'path' in a graph G is some sequence of edges where each edge begins with the vertex ending the preceding edge. That is, a path is any unbroken, non-intersecting route that may be traced along the edges in a graph. A path may be traversed through a graph either along or against

the arrows. If every edge in a path contains an arrow pointing from the first to the second vertex of the pair, the path is called a 'directed' path. When there is a path between two vertices in a graph the two vertices are labelled 'connected', and if not, they are labelled 'disconnected'. For three subsets a, b, and s of V, s is said to 'separate' a and b when all paths from a to b intersect s (Pearl 2000a:12-13).

A directed graph may include directed cycles, but directed graphs may not contain self-loops (eg. $X \Rightarrow X$). A graph that contains no directed cycles is called 'acyclic' and a graph that is both directed and acyclic is called a 'directed acyclic graph' (DAG). Graph theory makes use of kinship relations to refer to specific relationships amongst vertices (nodes) in a graph (eg. parents, children, descendants, and ancestors). Kinship relations are defined along the full arrows in the graph, including arrows that form directed cycles but ignoring bi-directed and undirected edges. For instance, a 'family' in a graph is a set of nodes made up from a specific node together with all its parents (Pearl 2000a: 12-13). A 'backdoor' path from one node X to another node Y is a path whose first edge is an arrow pointing into X. A 'blocked' path between nodes X and Y is a path that passes from a parent to a child and then on to another parent (Greenland and Brumback 2002: 1030-1031). A node in a directed graph is called a 'root' if it has no parents (explicitly represented as nodes in the graph) and a 'sink' if it has no children. Every DAG has at least one root and at least one sink. A connected DAG such that every node has at least one parent is called a 'tree'. A tree in which every node has at most one child is called a 'chain', and a graph in which every pair of nodes is connected by an edge is called 'complete' (Pearl 2000a: 12-13). Graphs typically summarize relationships between individuals within a population where each variable represents the states of individuals within that population. A population may contain just one individual and 'individuals' may represent almost any unit of interest to the modeller (Greenland and Brumback 2002: 1030-1031). A Bayesian Network has a function associated with each of its nodes – which is either a probability density function or a conditional probability table – stating that node's dependence upon its parents, or, when the node is a root, associating it with a prior probability distribution (Twardy and Korb 2002: 2).

Since they are a form of graph, each of these conditions and terminology applies to Bayesian Networks. Even so, Bayesian Networks perform a different role to that of graphs found in other forms of modelling such as regression and factor modelling[54]. The primary role of Bayesian Networks in statistical modelling is to provide guidance when drawing inferences under conditions of partial knowledge or uncertainty. In this sense Bayesian Networks are a graphical method for the representation and manipulation of probability distributions. As Kennett *et al.* (2001) explain, Bayesian Networks

> [...] allow a probability distribution to be decomposed into a set of local distributions. The network topology, and associated independence semantics, indicates how these local distributions should be combined to produce the joint distribution over all random variable nodes in the network (Kennett *et al.* 2001: 3)

And, since a statistical model is a type of probability distribution it follows that Bayesian Networks can represent the properties of statistical models. In practice many statistical models of interest contain enough variables so as to make the (stepwise) drawing of inferences a practical impossibility. In such cases Bayesian Networks function as data reduction and/or efficient organization devices. In fact, the primary practical purpose of a Bayesian Network includes the

> [...] substantial reduction in the number of parameters required to specify [a] distribution when the network connectivity is low; the ability to use more efficient algorithms for local distributions; and the separation of the quantification of influence strengths from the qualitative representation of the causal influences between variables [...] (Kennett *et al.* 2001: 3-4).

[54] See Murphy (2001) and Edwards (2000) for a useful overview.

Constructing a Bayesian Network has two parts – the specification of the structure of the network domain and the quantification of the dependencies and independencies so specified (Kennett *et al.* 2001: 4). Hence, the term 'structure' is to be identified at this stage with relations of probabilistic dependence and independence. The nature of these relationships as they pertain to Bayesian Networks is as follows. Granted a statistical parameter is the specification of a joint probability function over a distribution of variables, conditional independence may be defined in the following way. Given a finite set of variables $V = \{V_1, V_2, \ldots, V_n\}$ and a joint probability function P(.) defined over V, and where X, Y, Z are any three subsets of variables in V, X and Y are conditionally independent given Z if

$$P(x \mid y, z) = P(x \mid z) \text{ whenever } P(y, z) > 0. \quad (1.26)$$

That is, in the terminology of the 'information relevance' literature; 'learning the value of Y does not provide additional information about X, once we know Z'. Or, more commonly, Z 'screens off' X from Y. Pearl utilises the following notation, due to Dawid (1979), to denote the conditional independence of X and Y given Z[55]:

$$(X \perp\!\!\!\perp Y \mid Z)_P \text{ iff } P(x \mid y, z) = P(x \mid z)^{56} \quad (1.28)$$

Unconditional independence or marginal independence is then denoted by $(X \perp\!\!\!\perp Y \mid \varnothing)$; that is,

$$(X \perp\!\!\!\perp Y \mid \varnothing) \text{ iff } P(x \mid y) = P(x) \text{ whenever } P(y) > 0^{57}. \quad (1.29)$$

Conditional and marginal independence admit the following graphical analogues. Consider an imaginary data set consisting of a set of (arbitrarily ordered) measurements $\{V, W, \ldots, Z\}$ gathered on some population N. The measurements and

---

[55] Note that $(X \perp\!\!\!\perp Y \mid Z)$ implies the conditional independence of all pairs of variables $Vi \in X$ and $Vj \in Y$, but the converse is not necessarily true (Pearl 2000a: 11).

[56] I use the symbol '$\perp\!\!\!\perp$', for want of the appropriate symbol, to denote the screening-off relation.

[57] These definitions resemble those found in Pearl (2000a p 11) from where they are paraphrased.

the entities in N represent observations. When building a model of such data statisticians assume that the measurements are random variables with a joint probability function of the form:

$f_\theta(v, w, \ldots, z)$

where 'θ' is taken to be some undisclosed parameter. A statistical model then is a family of possible densities (such as $\{f_\theta : \theta \in \Theta\}$), where sub-models correspond to parameter subsets (Edwards 2000: 6). Given such a model, an undirected graph $G = (V, E)$ can be constructed such that V is the set of variables from the statistical model, and E represents the edges between pairs of variables that are not conditionally independent given the remainder of the variables in V. The general idea is to identify the statistical property of conditional independence with a graph-theoretic analogue. The analogue is labelled 'separation' (Edwards 2000: 7). With the identification of the graphical property of separation with the statistical property of independence it becomes possible to translate a set of conditional independence relations present in a statistical model into the language of Bayesian Networks. Pearl (2000a) describes the process in the following way.

Consider a distribution P defined on a distribution N of discrete variables ordered arbitrarily as $X_1, X_2, \ldots, X_n$. Using the chain rule of the probability calculus P may be decomposed as the product of n conditional distributions:

(PD) $P(x_1, \ldots, x_n) = \prod_j P(x_j \mid x_1, \ldots, x_{j-1})$.

It may turn out that the conditional probability of some variable X_j in N is independent of all its predecessors once the value of a select subset of predecessors is known. Call this subset PA_j and define the following equivalence:

(PDE) $P(x_j \mid x_1, \ldots, x_{j-1}) = P(x_j \mid pa_j)$.

The r.h.s of this equivalence may then be substituted into the product of (PD) in the process, simplifying the amount of information required to specify the probability of X_j. That is, given the decomposition scheme just outlined, less information is required to specify the probability of X_j conditional on just the possible realisations of the set PA_j rather than conditional on all possible realisations of X_j's predecessors $X_1,\ldots,X_{j-1}$ (Pearl 2000a: 14). Pearl labels the set PA_j the 'Markovian parents' of X_j, or 'parents' for short, and defines the set in the following way:

> *Markovian Parents*[58]
> Let $V=\{X_1,\ldots,X_n\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be the Markovian parents of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_j,\ldots,X_{j-1}\}$ satisfying $P(x_j \mid pa_j)=P(x_j \mid x_1,\ldots,x_{j-1})$ but such that no proper subset of PA_j does[59].

If it is the case that every configuration v of variables has some finite probability of occurring then the Bayesian Network associated with $P(v)$ is unique given the ordering of the variables[60] (Pearl 2000a: 15). According to the definition of Markovian parents Bayesian Networks are carriers of independence relationships along the order of construction. Hence every distribution that satisfies the definition of Markovian Parents must decompose into the product of

(PF) $P(x_1,\ldots,x_n)=\prod_i P(x_i \mid pa_i)$.

---

[58] This definition is from Pearl (2000a: 14).
[59] Lower-case symbols such as; 'x', 'paj' denote particular values of their corresponding variables written in upper-case symbols.
[60] See Pearl (1988) for a proof.

This product decomposition does not depend upon any specific variable ordering since, given the distribution *P* and its graph *G*, one can calculate whether *P* decomposes into the product given by (PF) without reference to the ordering of the variables in *P*. It therefore follows according to Pearl, that "... a necessary condition for a DAG *G* to be a Bayesian network of probability distribution *P* is for *P* to admit the product decomposition dictated by *G*" (Pearl 2000a: 16). Pearl captures this notion through the following definition:

Markov Compatibility[61]
If a probability function P admits the factorisation of (PF) relative to a DAG *G*, then *G* represents *P*, or, in other words, *P* and *G* are (Markov) compatible.

Pearl claims it is important to ascertain the compatibility between DAGs and probabilities in statistical modelling because compatibility is a necessary and sufficient condition for a DAG *G* to explain a body of empirical data represented by *P* (Pearl 2000a: 16).

Pearl offers a decision procedure for identifying conditional independencies in Bayesian Networks called 'd-separation'. The d-separation criterion is a rule-guided method for deciding, given three disjoint sets of variables $\{X\}$, $\{Y\}$, $\{Z\}$ represented by a graph *G*, whether *X* is independent of *Y* given *Z*. Put simply, d-separation offers a procedure for translating the dependence and independence relationships present in a graph into equivalent statements expressed solely in the language of probability. Pearl intends that the condition of statistical dependence should be associated with connectedness in the graph-theoretic sense, and that statistical independence should be associated with the absence of a connected path (i.e. with separation). The upshot is that the 'd' in 'd-separation' connotes directionality and the condition of d-separation and d-connection is intended to account for the direction of the arrows present in some DAG *G* and hence the dependence relations present in the statistical model. d-separation is defined as follows:

d-separation[62]

A path p is said to be 'd-separated' or 'blocked' by a set of nodes Z iff

1. p contains a chain i $\Rightarrow$ m $\Rightarrow$ j or a fork i $\Leftarrow$ m $\Rightarrow$ j such that the middle node m is in Z, or
2. p contains an inverted fork or collider i $\Rightarrow$ m $\Leftarrow$ j such that the middle node is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y iff Z blocks every path from a node in X to a node in Y.

Given two singleton variables x and y that form part of a DAG, the application of the criterion of d-separation obeys the following three rules[63]:

(Rule 1) Unconditional Separation

x and y are d-connected if there is an unblocked path between them.

An unblocked path is a path that can be traced without traversing a pair of arrows that collide 'head-to-head'[64].

(Rule 2) Blocking by Conditioning

x and y are d-connected, conditioned on a set Z of nodes, if there is a collider-free path between x and y that traverses no member of Z. If no such path exists, x and y are d-separated by Z or, in other words, every path between x and y is 'blocked' by Z.

(Rule 3) Conditioning on Colliders

---

[61] The definition of Markov Compatibility is from Pearl (2000a: 16).
[62] This definition is from Pearl (2000a: 16-17).
[63] These rules appear in Pearl (2001b).
[64] Pearl notes the ramification of rule 1 is that the covariance terms corresponding to these pairs of variables will be zero, for every choice of model parameters.

> If a collider is a member of the conditioning set Z, or has a descendant in Z, then it no longer blocks any path that traces this collider.

The application of these rules to a DAG allows the investigator to read off every d-separation relation contained in the graph and to infer that these relationships obtain in the distribution as statistical independencies if the model is correct.

The last condition I wish to introduce before moving on to the remaining components of the syntax to Pearl's account is that of 'observational equivalence'. This condition is used to determine whether every probability distribution that is compatible with one given DAG is compatible with another.

> *Observational equivalence*[65]
> Two DAGs are observationally equivalent if and only if they have the same skeletons and the same set of v-structures (i.e. two converging arrows whose tails are not connected by an arrow).

According to Pearl (2000a):

> Observational equivalence places a limit on our ability to infer directionality from probabilities alone. Two networks that are observationally equivalent cannot be distinguished without resorting to manipulative experimentation or temporal information (Pearl 2000a: 19-20).

Using the d-separation criterion it is possible to determine the set of independencies consistent with a specific DAG. It is also possible to specify all the DAGs implied by a given set of independencies. Using the criterion of observational equivalence it is then possible to determine which of these DAGs has the same set of v-structures. The fact that the set contains DAGs where the orientation of one or more arrows may be

[65] The definition of observational equivalence is provided by Theorem 1.2.8 from Pearl (2000a: 19).

reversed without thereby destroying any *v*-structures reinforces the point that Bayesian Networks encode associational information.

Although Bayesian Networks are the foundation of Pearl's account, they do not take on any lustre until Pearl provides them with a causal interpretation. After outlining the causal interpretation of Bayesian Networks Pearl introduces the causal Markov condition and thus moves on to the explication of functional causal models. I will not detail the causal interpretation of Bayesian Networks here. The appropriate place to do so is in section 2.0.3 where I discuss the account's informal semantics. The reason is straightforward; strictly speaking, the account is at this point un-interpreted formalism. The idea that Bayesian Networks may encode causal relations is read into their use as models. I discuss this issue at length in the following chapter. Suffice it to say that the general idea of interpreting Bayesian Networks as causal diagrams involves assuming each parent-child relationship in a Network to represent an 'autonomous mechanism' whereby such mechanisms underpin or produce the conditional independence relations manifested by a network (Pearl 2000a: 22-23).

Getting back to the formalism, Pearl defines a causal Bayesian Network in the following way:

> *Causal Bayesian Networks*[66]
> Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x. Denote by P^* the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$, which represents no intervention (i.e. $X = \varnothing$). A DAG G is said to be a causal Bayesian network compatible with P^* if and only if the following three conditions hold for every $P_x \in P^*$:
>
> (i) $P_x(v)$ is Markov relative to G;

[66] The definition of causal Bayesian networks is from Pearl (2000a: 23).

(ii) $P_x(v_i) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$;

(iii) $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$.

Pearl contends that whenever a graph G is a causal Bayesian Network with respect to P^* the following two properties of G can be proven:

Property 1:

For all i,

$P(v_i \mid pa_i) = P_{pa_i}(v_i)$

Property 2:

For all i and for every subset S of variables disjoint of $\{V_i, PA_i\}$:

$P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$.

According to Pearl, Property 1 renders every parent set (PA_i) exogenous relative to its child V_i, thus ensuring that the conditional probability $P(v_i \mid pa_i)$ coincides with the effect (on V_i) of setting PA_i to pa_i by external control. Property 2 states that, once we control its direct causes PA_i, no other interventions will affect the probability of V_i (Pearl 2000a: 21-24).

The next step Pearl takes in setting out the syntactic machinery of his account is to generalise the machinery of the causal Bayesian Network to 'functional causal models' and to introduce the 'causal Markov condition'. The value of generalising causal Bayesian Networks to functional causal models lies in the wider applicability of functional causal models to problems involving the quantification of causal relationships and the fact that functional causal models make the structural component of graphs more perspicuous. A functional causal model consists of some number of equations each possessing the following form:

$x_i = f_i(pa_i, u_i),\ i = 1,\ldots,n,$

where 'pa_i' stands for specific values of variables from the set of variables judged to be 'immediate causes' of x_i, and where 'u_i' stands for specific values from the set of the disturbances or errors that result from factors omitted from the model. Pearl calls a set of such functional equations (and interprets each equation to represent an 'autonomous mechanism') a 'structural model'. When, in any given structural model, each equation (mechanism) determines the value of only one variable the model is referred to as a 'structural causal model' and the sole variable is labelled the 'dependent' variable (Pearl 2000a: 27). Pearl demonstrates that structural causal models are capable of representing the same structures of dependence and independence as Bayesian Networks. The first step in the demonstration is to introduce the notion of a 'causal diagram'.

Given a structural causal model, its accompanying graph *G*, called a 'causal diagram', may be constructed by drawing an arrow from each member of the set PA_i toward the set X_i. If the causal diagram that results from this procedure is acyclic the corresponding structural causal model is labelled 'semi-Markovian' and each of the *X* variables will be uniquely determined by the *U* variables. If the disturbance terms of this acyclic causal diagram are mutually independent then the corresponding structural causal model is labelled 'Markovian' and the following theorem holds:

> *Causal Markov Condition*[67]
> Every Markovian causal model M induces a distribution $P(x_1,\ldots,x_n)$ that satisfies the parental Markov condition relative to the causal diagram G associated with M; that is, each variable X_i is independent on all its non-descendants, given its parents PA_i in G.

[67] The definition of the Causal Markov Condition is from Pearl (2000a: 30).

For Pearl the causal Markov condition shows that each child-parent relationship in a model can be read as both a deterministic function and a conditional probability. In fact Pearl takes the two to be equivalent in-as-much as the functional reading imposes equivalent independence constraints on the resulting distribution and lead to the same recursive decomposition that is characteristic of the probabilistic reading. The upshot then is that

> [...] for every Bayesian network *G* characterised by a distribution *P* (as in (PF) above), there exists [at least one] functional model that generates a distribution identical to *P*. It follows that in all probabilistic applications of Bayesian networks [...] [the investigator] can use an equivalent functional model [...] and [...] can regard functional models as just another way of encoding joint probability functions (Pearl 2000a: 30-31).

The last major component of Pearl's syntactic machinery is the notational norms and inference rules that govern what and how interventions or surgeries may be performed upon models and what form the results of such interventions must take. Pearl labels his notation and inference rules the 'calculus of interventions' or the 'do-calculus' for short[68]. The overall point of the calculus is to provide a (syntactic) method by which sentences concerning interventions may be transformed into equivalent observational sentences that pertain to post-intervention effects (Pearl 2000a: 85).

Define the simplest type of intervention as one in which a single variable X_i is fixed to a constant value x_i, which in the process destroys the pre-intervention influence of the old functional mechanism and installs a new functional mechanism in its place. Call such an intervention 'atomic'. In Pearl's words, an atomic intervention "... amounts to lifting X_i from the influence of the old functional mechanism $x_i = f_i(pa_i, u_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed" (Pearl 2000a: 70). The notation

used to denote an atomic intervention is $do(X_i = x_i)$, or $do(x_i)$ for short. Performing an intervention creates a new model (and graph) that represents what happens to the original model when it is manipulated. That is,

> [...] when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the [causal model], one for each member of X, thus defining a new distribution over the remaining variables that completely characterises the effect of the intervention (Pearl 2000a: 70).

When used to alter graphs the do-calculus obeys the following set of inference rules. Where X, Y, and Z are arbitrary disjoint sets in a DAG G, denote by '$G_{\overline{X}}$' the graph that results from deleting all arrows in G pointing to nodes in X. Denote by '$G_{\underline{X}}$'the graph that obtains by deleting all arrows that emerge from nodes in X. Denote by '$G_{\overline{X}\underline{Z}}$' the graph that results from the deletion of both incoming and outgoing arrows in G. Finally the expression $P(y \mid \hat{x},z) \cong P(y,z \mid \hat{x})/P(z \mid \hat{x})$ stands for the probability of $Y = y$ given that X is held constant at x and that (under this condition) $Z = z$ is observed (Pearl 2000a: 85).

The basic inference rules of the do-calculus are encapsulated in the following three rules[69]:

> *Inference Rules of do-calculus*
> Let G be the directed acyclic graph associated with a causal model defined in (3.2), and let $P(.)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W, we have the following rules.

---

[68] This has been called the 'set(x) calculus' elsewhere.
[69] The proofs of these rules appear in Pearl (1995) and this presentation of the rules is from Pearl (2000a: 85-86).

Rule 1 (Insertion/deletion of observations):

(3.31) $P(y \mid \hat{x},z,w) = P(y \mid \hat{x},w)$ if $(Y \amalg Z \mid X,W)_{G_{\bar{X}}}$

Rule 2 (Action/observation change):

(3.32) $P(y \mid \hat{x},\hat{z},w) = P(y \mid \hat{x},z,w)$ if $(Y \amalg Z \mid X,W)_{G_{\bar{X}\underline{Z}}}$

Rule 3 (Insertion/deletion of actions)

(3.33) $P(y \mid \hat{x},\hat{z},w) = P(y \mid \hat{x},w)$ if $(Y \amalg Z \mid X,W)_{G_{\bar{X}\bar{Z}(\bar{W})}}$ where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\bar{X}}$

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $do(X = x)$, hence the graph $G_{\bar{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms. Rule 2 provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all backdoor paths[70] from Z to Y (in $G_{\bar{X}}$), since $G_{\bar{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$)[71] (Pearl 2000a: 85-86).

2.0.2 Formal semantics

[70] Where a set of variables Z satisfies the 'backdoor' criterion relative to an ordered pair of variables (Xi, Xj) in a DAG G just in case: (i) no node in Z is a descendant of Xi, and (ii) Z blocks every path between Xi and Xj that contains an arrow into Xi. See Pearl (2000a: 79-80) for further details and discussion.

[71] Proofs of the inference rules of the do-calculus are presented in Pearl (1995).

The move from Bayesian Networks to structural causal models accompanied by the calculus of intervention culminates in the definition of 'causal model'. Once causal models and their properties are defined Pearl introduces a further two objects to complete the formal component of the account–the remainder being left to interpretation. The first of these objects is the 'causal world' and the second is the 'causal theory'. The introduction of causal models, worlds and theories marks the shift in Pearl's account from syntax to semantics (Pearl 2000a: 202). I begin with Pearl's definition of causal models[72].

Causal Model[73]

A causal model is a triple $\langle U,V,F \rangle$, where:

(i) U is a set of background or exogenous variables that are determined by factors outside the model.

(ii) V is a set of variables $\{V_1,V_2,\ldots,V_n\}$ labelled endogenous that are determined by variables in $U \cup V$.

(iii) F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from the respective domains of $U \cup (V \setminus V_i)$ to V_i and such that the entire set F forms a mapping from U to V. That is, each f_i defines the value of V_i given the values of all other variables in $U \cup V$, and the entire set F has a unique solution $V(u)$. The members of the set of F can be represented by $v_i = f(pa_i,u_i)$, $i=1,\ldots,n$, where 'pa_i' is any realisation of the unique minimal set of variables PA_i in V\Vi sufficient for representing f_i. Likewise, '$U_i \cup U$'

[72] It is important to note that Pearl's account was developed over a number of years in conjunction with broader developments in the fields of statistics and Artificial Intelligence and is still undergoing a process of extension and refinement. There are several variations of the account's notational and formal (syntactic) structure and formal semantics available in the literature. Even so, I take it that the key components of the semantics are the interpretation provided to the calculus of interventions and the evaluation of causal sentences.

stands for the unique minimal set of variables in U sufficient for representing f_i[74].

All causal models can be associated with a graph, or what Pearl now calls a causal diagram, $G(M)$ such that each node of the graph G corresponds with a variable of the model and the directed edges point from a member of the sets PA_i and U_i towards a member of V_i (Pearl 2000a: 203). Once provided with a set of measurements and some theoretical context, the model is taken to represent a portion of reality. In other words, a causal model makes a number of positive assertions about the nature of some portion of the world where such assertions are true conditional on the efficacy of the model (Pearl 2000a: 202). This relationship holds also in the case of a 'probabilistic causal model' except that in the case of the latter the correspondence relation is mediated via the investigator's beliefs about the likely truth of the model given available evidence.

Probabilistic Causal Model
A probabilistic causal model is a pair

$$\langle M, P(u) \rangle,$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

The most important feature of causal models is their capacity to represent the effects of interventions. As with causal Bayesian Networks, interventions can involve a simple action such as setting a variable to a constant or some more elaborate action involving the quantification that a change in one variable has on another.

---

[73] The definition of causal model is from Pearl (2000a: 203). But see also the version presented below belonging to Menzies (2002), which is adapted from the account provided by Halpern and Pearl (2002a).

Interventions on causal models amount to transforming a pre-intervention model into a post-intervention model. Pearl calls post-intervention models 'submodels' and provides them with the following definition:

> Definition 7.1.2 (Submodel)[75]
> Let M be a causal model, X a set of variables in V, and x a particular realization of X. A submodel M_X of M is the causal model
>
> $$M_X = \langle U, V, F_X \rangle$$
>
> Where
>
> $$F_X = \{f_I : V_i \notin X\} \cup \{X = x\}$$

What this means is that a submodel F_X is formed by deleting from a causal model F all functions F_i that correspond to the set of variables X and replacing them with the set of constant functions $X = x$ (Pearl 2000a: 204). This idea is encoded in the following definition:

> Definition 7.1.3 (Effect of Action)[76]
> Let M be a causal model, X be a set of variables in V, and x a particular realization of X.
> The effect of action $do(X = x)$ on M is given by the submodel M_X.

The idea here is that the transformation of a causal model M to a submodel M_X results in an alteration of the content of the set of model functions F. The transformation opens the possibility of using the set of functions F_X to calculate the

---

[74] Pearl defines sufficiency in this setting in the following way. A set of variables X is sufficient for representing a function y = f(x, z) if f is 'trivial' in Z (that is, if for every x, z, z' we have f(x, z) = f(x, z').

[75] This definition is from Pearl (2000a: 204).

[76] This definition is from Pearl (2000a: 204).

value that variables besides X assume in response to actions. Pearl labels this the 'potential response' a variable Y has to an intervention performed on another variable X:

> Definition 7.1.4 (Potential Response)[77]
> Let X and Y be two subsets of variables in V. The potential response of Y to action $do(X = x)$, denoted $Y_X(u)$, is the solution for Y of the set of equations F_X.

Likewise, submodels may be used to determine the value that a variable would have obtained had a distinct variable of the model undergone a 'natural change' or been intentionally manipulated. Pearl sees such relations as counterfactual and thinks of interventions in such cases as hypothetical modifications of the model's equations aimed at simulating what would have happened in reality had nature's mechanisms been (minimally) altered (Pearl 2000a: 205). Such counterfactual phrases are defined as a form of potential response function:

> Definition 7.1.5 (Counterfactual)
> Let X and Y be two subsets of variables in V. The counterfactual sentence 'The value that Y would have obtained, had X been x' is interpreted as denoting the potential response $Y_X(u)$.

In fact, Pearl believes that this definition provides the functions of a causal model with interventional sense precisely because $v_i = f_i(pa_i, u_i)$ is just the value of V_i in the submodel $Mv \setminus v_i$. That is, "[...] $f_i(pa_i, u_i)$ stands for the potential response of V_i when we hold constant *all* other variables in V" (Pearl 2000a: 205 original emphasis).

Pearl provides the following procedure – summarised by a theorem – for calculating the values of counterfactual quantities in probabilistic causal models:

Theorem 7.1.7[78]

Given a model $\langle M, P(u)\rangle$, the conditional probability $P(B_A \mid e)$ of a counterfactual sentence 'If it were A, then B', given evidence e, can be evaluated using the following three steps.

1. *Abduction* – Update $P(u)$ by the evidence e to obtain $P(u \mid e)$.

2. *Action* – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.

2. *Prediction* – Use the modified model $\langle M_A, P(u \mid e)\rangle$ to compute the probability of B, the consequent of the counterfactual.

The definition of causal models can be extended to include 'causal worlds' and 'causal theories'. Pearl defines these objects as follows:

Worlds and Theories[79]

A causal world w is a pair $\langle M, u\rangle$, where M is a causal model and u is a particular realisation of the background variables in U. A causal theory is a set of causal worlds.

Pearl asserts that a world w can be viewed as a probabilistic model for which $P(u) = 1$. Causal theories are used in Pearl's account to characterise partial specifications of causal models. That is, for example, with models sharing the same causal diagram or models in which the 'f_i' are linear with undetermined coefficients (Pearl 2000a: 205-207). However, even though Pearl compares his account of partially specified models with Lewis's account of possible worlds, Pearl does not

[77] This definition is from Pearl (2000a: 204).
[78] This theorem is from Pearl (2000a: 206).
[79] This definition is from Pearl (2000a: 207).

seem to be aware of Lewis's modal realist commitments[80]. The two accounts cannot, therefore, be straightforwardly equated.

Given the specificity of the model, truth-values may be assigned to sentences that take, or may be transformed into those of the following three generic forms:

> *Predictions* – "The value of y would be n, given we know x."
> *Interventions* – "The value of y would be n if we set the value of x to n."
> *Counterfactuals* – "The value of y would have been n if the value of x was n, even given that the actual value of y is n' and x is n''."

According to Pearl (2000a), providing truth-values to each of these three generic sentences requires an increase in the level of specific information needed to arrive at a decision as one moves from predictive sentences through to counterfactual sentences. Pearl comments that evaluating predictive sentences is the simplest of the three tasks because it requires only the specification of a joint probability function. Interventions are more involved since they require the specification of a joint probability function and sufficient information about the causal structure of the phenomena under investigation. Counterfactuals are more labour intensive again since their evaluation requires some information concerning the functional relationships between variables if not also information pertaining to the distribution of the omitted factors that make up the model (Pearl 2000a: 38). Sentences taking one or another of these forms are considered to be generic in the sense that the forms offer a 'well-formed' sentential vehicle with which to pose questions to a given model. All that is required to complete any one of these sentence forms is to substitute the sentence's variables for values. Then, if the resulting sentence follows from a model of interest that sentence is true. In effect, a causal model entails the truth of a sentence just in case that sentence is either trivially true given the model, or the sentence can serve as the conclusion of a valid argument of which the (sub-)model serves as a set

[80] This may not affect the comparison however, since Lewis also thought that the relevant counterfactual truths about the actual world are true solely in virtue of features of this world. See

of premises, or the sentence may be deduced from a model via the appropriate inferences rules[81]. The idea is extended to causal quantities through the specification of criteria for identifiability[82].

2.0.3 Informal Semantics

As one would expect, the formal semantics is accompanied by an informal treatment[83]. In a general sense an informal treatment of the formal semantics is necessary so as to add meaning to the (principle terms of the) latter since, without an informal treatment, the formal semantics remains un-interpreted. The trouble with leaving the semantics un-interpreted in this instance is that the formal semantics does not employ the terms cause, effect, truth and so forth in a meaningful way. Left in such as state the formal semantics is not a *causal* semantics and the models consistent with the semantics do not say anything about causal relationships[84]. The principal goal of the informal semantics is in this instance to provide 'X is the cause of Y' and other generic expressions of the formal semantics with a meaningful or intended interpretation given the context within which such sentences appear'[85]. Furthermore, since the notion of truth that Pearl uses attaches to models, which are intended to be models of reality, that notion will in turn depend for its meaning on what Pearl has to say about how models represent reality. It is interesting to note that a large portion of Pearl (2000a) is spent interpreting the account's semantics and yet one cannot readily locate the kernel of the account[86]. Here I set out only part of what I take to be the

---

Armstrong (1999b) for discussion.

[81] Completeness proofs are outlined in Pearl (2000a: 230-231).

[82] The notion of identifiability for such models involves determining the quantitative effect one variable of a model has on another. See Pearl (2000a: 91-96) for discussion. I examine related issues in section 3.4 of the following chapter.

[83] Principally canvassed in Section 7.2 p 215.

[84] For details of the difference between a formal and an informal semantics see for instance Copeland (1983), Plantinga (1974) and Haack (1978: 188-190).

[85] Berkovitz (2002) offers an argument to this effect. I outline his argument in the following chapter.

[86] Pearl is criticised on this point by several philosophers who assert that Pearl (2000a) says little about what causality is. I take this issue up in detail in chapter 3. But, see Hitchcock's comments in his (2001).

informal interpretation before taking up further issues for discussion in later sections as the need arises. The most important components of the informal account to be placed on the table at this point are, to my mind, Pearl's interpretation of his counterfactual semantics; his account of types versus token causes; and of the mechanisms that underlie a causal model's functions. The sense Pearl (2000a) provides to each of these is somewhat intertwined. I begin with Pearl's interpretation of functions and follow with mechanisms and close with Pearl's interpretation of type and token causes.

It is clear from the discussion above that Pearl (2000a) thinks mechanisms should be equated with a specific form of mathematical function. What is not obvious at first glance is how this function is to be understood. Two questions need to be decided. First, what is the mathematical form of the function? Second, what interpretation does Pearl provide such forms? Both questions can be addressed together.

At the outset Pearl asks that the reader take note of the difference between the use of the '=' sign to represent equality in algebraic systems and its meaning within Pearl's formalism. Pearl states that in his formalism the '=' acts like an assignment operator in the sense that the variable appearing on its l.h.s. is assigned a value according to the solution of the terms appearing on the '=' sign's r.h.s. but not vice versa. There is a related difference to be noted with the interpretation of each equation. In Pearl's words:

> A set of mechanisms, each represented by an equation, is not equivalent to the set of algebraic equations that can be assembled from those mechanisms. Mathematically, the latter is defined as *one* set of n equations, whereas the former is defined as n separate sets, each containing one equation. These are two distinct mathematical objects that admit two distinct types of solution-preserving operations. The calculus of causality deals with the dynamics of such modular systems of equations, whereas the addition and deletion of equations represent interventions (Pearl 2002b: 212 original emphasis).

The latter portion of this comment implicates the distinction Pearl draws between 'seeing' and 'doing' as a component in the interpretation of model functions. The distinction is relevant in the sense that the notion of 'doing' coincides with the 'equation wipe-out' interpretation of intervention and an intervention on a variable within a causal model is represented as the deletion of one equation and its replacement with another. That is, interventions 'wipe-out' an equation and in so doing place some relevant variable under the influence of a new function. How interventions may be performed is governed by the rules of the do-calculus. The reason that such a 'calculus of interventions' is necessary, Pearl asserts, is that interventions cannot be successfully expressed using the probability calculus alone. Probability theory is inadequate, according to Pearl, because:

> ...probability theory deals with beliefs about an uncertain, yet static world, while causality deals with changes that occur in the world itself. Causality deals with how probability functions change in response to new conditions and interventions that originate from outside the probability space, while probability theory, even when given a fully specified joint density function on all variables in the space, cannot tell us how that function would change under external interventions. Thus, "doing" is not reducible to "seeing", and there is no point trying to fuse the two together. [...] The additional information needed for making [predictions about change] is analogous to the causal information (about invariant mechanisms) that the do calculus extracts from a directed acyclic graph (DAG) (2002a: 208)[87].

In turn notions of intervention and of 'doing' are underpinned by the notion of 'modularity'. Modularity is, for Pearl, an assumption pertaining to mechanisms such that each parent-child relationship represented by a graph may be changed without

[87] This quote illustrates well Pearl's strong commitment to a subjective interpretation of probability. It almost goes without saying that interpretations of probability are always contentious in one way or another. Many scientists trained to think of probabilities as frequencies or one sort or another will find Pearl's interpretation especially contentious.

thereby affecting changes to any other such relationships in the graph. This means that the overall effect of an intervention can be predicted by changing the relevant equations in a model and then using the modified model to compute the state of the new–post-intervention–model that results (typically a probability function) (Pearl 2000 p 32)[88]. These views mark a break with the common or modern interpretation of model structure found in, for example, Duncan (1975), Kline (1998), Blalock (1985) and Bollen (1989). Pearl intends to re-capture the interpretation given to model structure by Wright (1921), Haavelmo (1943) and Koopmans (1950, 1985). What the functions of Pearl's formalism do is relate the values of the set of variables, which are 'parents' of a variable x, with a proportion that affects x from outside the model and equates the result with the value of x. In other words, x's value is a function of x's 'direct causes' and the invariant properties summarised by a model's u's. Pearl offers the following insight into how these functions are to be interpreted:

> [...] while I was working with Tom Verma on "A theory of Inferred Causation" [we] played around with the possibility of replacing the parents-child relationship $P(x_i \mid pa_i)$ with its functional counterpart $x_i = f_i(pa_i, u_i)$ and, suddenly, everything began to fall into place: we finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities $P(x_i \mid pa_i)$ with which we had been working so long in the study of Bayesian Networks (Pearl 2000a: 104).

The specific mathematical form of such functions depends upon the application to which the model is put since the nature of the variables and the expression afforded to variables in the model varies from application to application. In models of the social and behavioural sciences, for instance, the functions are assumed to be linear. Even so, it is generally accepted since Spirtes (1994) that the form of the functions of a causal model is arbitrary and that it is more important to recognise the independence

[88] However, it is important to note that this works at the level of the model in contrast to what is being modelled. It is then a conceptual definition and not a claim about model validity or verification. See

of the background variables associated with each equation of the model. (Scheines 1997: 192; Pearl 2000: 69)[89].

I now move on to the details of Pearl's interpretation of counterfactuals. The reader will recall from above that Pearl ties the syntactic account of interventions to counterfactuals in the formal semantics. Pearl remarks that his semantics for counterfactuals is closely related to those set out in Lewis (1973a, 1973b, 1979). Pearl's interpretation of counterfactuals is, however, considerably different. As this is so, a brief introduction to the nature of counterfactual conditionals is in order.

'Counterfactuals' can be thought of as a species of conditional. In the context of natural language utterances, a conditional utterance, broadly construed is one containing an 'if' clause such as 'If Bob is at the shop, Wendy is at the shop'. A counterfactual conditional utterance can then (though not universally) be heard as a conditional utterance in the subjunctive mood. An example is the following: 'If Bob was at the shop, Wendy would have been at the shop.' There are numerous theories of conditionals. Some draw an explicit distinction between counterfactual conditionals and other conditional utterances. I have introduced counterfactuals by way of utterances only to flag the fact that the following discussion of Pearl's view of counterfactuals takes the expressions and inferences of scientists as its starting point. It is far too great a project to attempt a summary of available theories of conditional utterances here. Suffice it to say that debate continues about the correct way to analyse both conditional and counterfactual utterances. By and large, philosophers' interest in conditionals and counterfactuals has tended to be with semantic properties. In turn, the interest in semantic properties has tended to centre on investigating the truth conditions for conditionals and counterfactuals[90].

---

further discussion in chapter 3 below.

[89] See also the discussion of identifying causal effects in Pearl (2000a: 92-93).

[90] See for instance Edgington (1995), Adams (1975), Lycan (2001), Lewis (1973b), Jackson (1979, 1991), Jeffrey (1964), Horwich (1990), Kvart (1986) Barker (1995). Though some, such as Adams, deny that (indicative) conditionals have truth conditions.

Counterfactuals have also been implicated in the task of providing an account of causation. A basic version of a counterfactual analysis of causation takes the following form:

> Where C is the putative cause and E the effect, the proposition 'C causes E' is true if and only if it is true that 'If C had not occurred then E would not have occurred'[91].

The broad idea behind a counterfactual analysis of causation is that (the truth conditions of) causal claims are reducible to counterfactual claims. It is common therefore to see the task of detailing a counterfactual account of causation to be that of showing all (relevant) causal utterances (or propositions) to be counterfactual utterances (or propositions). Here it is important to note the way in which an analysis of counterfactual utterances differs from an analysis of counterfactual propositions. The main difference between the two is that the former focuses on actual linguistic practices of language users as its object of study whilst the latter focuses on accounting for the truth-conditions of propositions. The two are, however, obviously related. Pearl, for example, briefly discusses the meaningfulness of counterfactual utterances commonly made by scientists before going on to assimilate such utterances under the umbrella of a small set of 'generic' counterfactual sentences. Pearl then proceeds to account for the truth-conditions of such sentences. Suffice it to say that, in the present context, it is typically the latter task that attracts the most attention.

It is standard to argue that providing truth conditions for counterfactuals requires one to consider states of affairs that do not actually transpire. That is, to consider what 'would occur' or 'would have occurred' but in actuality does not. Many have objected to counterfactuals on the grounds that what does not occur cannot be observed and since counterfactuals, by definition, do not occur and so cannot be observed, counterfactuals are not verifiable. The objection amounts to the claim that

---

[91] Throughout the literature C and E have been rendered as either event *types* or *singular* events. Pearl's account of type and token causes is at variance with the literature on this point.

counterfactual propositions have no empirical content and so cannot be 'tested' or empirically verified.

For instance, Dawid (2000) asserts in the spirit of Popper that

> [...] the meaningfulness of a purportedly scientific theory, proposition, quantity, or concept is related to the implications it has for what is or could be observed, and, in particular, to the extent to which it is possible to conceive of data that would be affected by the truth of the proposition or the value of the quantity. When this is the case, assertions are empirically refutable and are considered 'scientific.' When this is not so, they may be branded 'metaphysical.' I argue that counterfactual theories are essentially metaphysical (Dawid 2000: 408)[92].

It is against this sort of view of counterfactuals that Pearl (2000) is reacting. According to Pearl (2000a) counterfactuals do not stand contrary to fact and can be empirically verified[93]. To illustrate its shortcomings Pearl (2000a) compares counterfactual propositions to scientifically accepted types of predictive propositions that are consistent with empirical laws. Pearl accepts that any number of propositions that make a prediction about the future states of a law governed system may be assessed for truth or falsity according to the laws that govern that system and measurements taken on the states of the system. But, in line with sentiments expressed by Dawid (2000) Pearl further accepts there is a *prima facie* problem with assessing counterfactuals precisely because such propositions appear to speculate about events that have not and could not have occurred and that consequently do not appear open to measurement let alone evaluation in the actual world (Pearl 2000a:

[92] Dawid (2000) does admit the use of counterfactuals for the purposes of causal modelling but denies that any inferences may be drawn from the counterfactual components of such models. See Dawid (2000: 409, 415-417) for discussion.

[93] I note in passing that when making their respective arguments regarding verification and falsification neither party appears aware of issues regarding confirmational holism discussed by Duhem (1954), Quine (1980, 1995) and Lakatos (1970).

217-218). But according to Pearl, in contrast with Dawid (2000), the problem is only apparent.

Pearl (2000a) sidesteps the difficulty of assessing counterfactuals by interpreting the counterfactual form as shorthand for a predictive form of conditional proposition. The key difference between the two is, according to Pearl, that the factual portion of the counterfactual form makes superfluous many specifications that are either left open or covered by a *ceteris paribus* clause in the predictive form. Hence,

> ...a counterfactual statement might well be interpreted as conveying a set of predictions under a well-defined set of conditions – those prevailing in the factual portion of the statement (Pearl 2000a: 219).

Since, for Pearl, counterfactual statements are a form of prediction, counterfactuals may be evaluated for truth and falsity in the same fashion as are predictions provided the following clause is respected. To successfully carry out an evaluation of a counterfactual Pearl (2000a) believes that the mechanisms and the boundary conditions of the model to which the counterfactual belongs must remain invariant. In Pearl's words:

> Cast in the language of structural models, the [mechanisms] correspond to the equations $\{f\}$ and the boundary conditions correspond to the state of the background variables U. Thus, a precondition for the validity of predictive interpretation of a counterfactual statement is the assumption that U will not change when our predictive claim is to be applied or tested (Pearl 2000a: 219)[94].

What to make of this view of counterfactuals? It is noteworthy that Pearl (2000a) does not draw any explicit distinction between so-called 'might', and 'would'

counterfactuals. In fact, Pearl (2000a) contains no discussion of different counterfactual (or conditional) moods whatever. Nor does Pearl discuss the syntactic structure of counterfactuals. This is surprising given Pearl's assertions about the content of counterfactual utterances, and especially so when the immediate context of such utterances is communication between scientists when reasoning about the causal relationships of some specific causal systems. It is also noteworthy that Pearl's standard for testability is not obviously applicable to counterfactuals that have a probabilistic component. In fact, Pearl's view of counterfactuals easily leads one to confusion in cases where Pearl appears to offer both conditional and subjunctive paraphrases of the same 'counterfactual' sentence[95].

The latter point is revealing. Pearl's assertion that counterfactuals may be interpreted as a set of predictions indicates that Pearl takes counterfactual conditionals to be of the same species as the so-called 'forward looking' indicative conditionals[96]. The key to the assimilation is clear from Pearl's assertion that predictive (forward looking indicatives) and counterfactual conditionals share a set of conditions which, when it can be ascertained that such conditions are invariant, guarantee that consequent follows from antecedent[97].

A closer look at how Pearl's account of counterfactuals differs from the account offered by Lewis is also instructive. There have been several attempts to specify the semantics of natural language counterfactual utterances. One prominent account is the so-called Stalnaker-Lewis account. It is generally assumed that natural language counterfactuals have the same syntactic structure as common conditional utterances[98]. As I mentioned above, the distinguishing feature of counterfactual conditionals is that

---

[94] Note that Pearl says 'on the assumption' and not 'with the knowledge or belief that' when speaking of invariance. This suggests that Pearl is thinking of invariance as an objective property of the system under investigation.

[95] See, for instance the comments and paraphrasing of such 'counterfactuals' by Pearl (2000a: 208).

[96] See Gibbard (1981) for comment on forward looking indicative conditionals and Lycan (2001) for criticism of the indicative/subjunctive classification of conditionals.

[97] It is not clear whether Pearl takes all counterfactuals in the context of scientific practice to be translatable into such indicative conditionals (though I suspect this to be the case).

the antecedent of the conditional is made true relative to some state or other that is not necessarily the one witnessed or expected. On the Stalnaker-Lewis account sorting out how to assign truth-values to such apparently non-actual states or events calls for the comparison of propositions across some number of possible worlds[99]. To carry out the comparison requires evaluating the truth or falsity of counterfactual conditionals according to a suitable ordering relation across a class of possible worlds. Lewis's specification of an ordering relation takes the following form:

> $\phi \,\Box\!\!\rightarrow \psi$ is true at a world *i* (according to a system of spheres **S**) if and only if:
>
> 1. no ϕ-world belongs to any sphere *S* in **S**, or
> 2. some sphere *S* in **S** contains at least one ϕ-world, and $\phi \supset \psi$ holds at every world in *S*[100].

where '$\phi \,\Box\!\!\rightarrow \psi$' represents the counterfactual conditional and where a sphere is made up of some number of possible worlds that are equally similar to the world at the centre of a system of spheres such that the larger the sphere the less similar are its worlds to the centring world[101]. In effect, Lewis's account of counterfactual conditionals offers a procedure whereby counterfactuals can be ranked according to how similar they are to what is true of some world at the centre of a system of spheres of increasingly different possible worlds. Hence, counterfactuals such as 'If P were the case then Q would be the case' are true of a possible world W iff there is some possible world W*, in which P and Q are true, and W* is closer to W than any world where P is true and Q is false[102]. Similarly, counterfactuals such as 'If P were the case then Q might be the case' are true in a possible world W iff there is some possible

---

[98] Though some think this may turn out to be a costly oversight. See, for instance, Lycan (2001) and Barker (1991).
[99] Where, for Lewis, possible worlds exist as concrete but non-actual entities.
[100] Reproduced from Tooley (2003: 371). See also Lewis (1973b: 16) and Lewis (1973a).
[101] Though Lewis and Stalnaker disagree over the nature of a similarity measure with a centring principle.
[102] Or where there is no world in which P is true.

world W*, in which P and Q are true, and W* is closer to W than at least one world where P is true and Q is false[103] (Tooley 2003: 372).

Pearl agrees that a measure of similarity amongst possible worlds is at the heart of Lewis's account, but is quick to mention his reservations as to whether a similarity measure of inter-world distances is the correct measure for evaluating counterfactuals (Pearl 2000a: 239). Pearl asserts that:

> In contrast with Lewis's theory, counterfactuals [on Pearl's interpretation] are not based on an abstract notion of similarity among hypothetical worlds; instead, they rest directly on the mechanisms that produce those worlds and on the invariant properties of those mechanisms (Pearl 2000a: 239)[104].

But, even so, Pearl later establishes via axomatic comparison that the formal component of his account is equivalent to Lewis's. The salient point is Pearl apparently agrees with Lewis on what the correct logic for reasoning with counterfactuals is but disagrees with Lewis on the correct interpretation of that logic[105]. Consequently, Pearl rejects Lewis's similarity measure and replaces it with a distance measure according to which the distance between two worlds $d(\omega,\omega')$ is the minimal number of local interventions (governed by the *do*-calculus) required to transform one world into the other (Pearl 2000a: 241). Modality is therefore defined relative to a causal model (or world)[106]. No wonder then that Pearl thinks the term 'counterfactual' is a misnomer. In effect Pearl is reading counterfactuals as a type of

---

[103] I leave aside the various amendments Lewis made to this account in his attempt to overcome its shortcomings. Nothing hangs on them for purposes of the present discussion.

[104] Note that Pearl considers Lewis to have adopted a 'hypothetical' view of possible worlds rather than the modal realist position Lewis actually held. This travels some way towards explaining why Pearl believes his account of counterfactuals does not differ greatly from Lewis's. Note also that Pearl suspects Lewis's account of counterfactual causation is circular since the similarity measure requires reference to causal laws. The claim that similarity judgements rest on causal judgements has been argued by Kitcher (1989).

[105] Though, I note that there are a number of questions left hanging about the details of this conditional logic. For instance, what theorems are valid in this logic and how exactly one is to paraphrase 'causal queries' into the formulae of this logic. For discussion of such issues see Priest (2001), Hopkins and Pearl (2003) and Meheus (2002) especially the article by Nickles.

indicative conditional sentence excepting that, where indicative conditional sentences in the present context are always accompanied by a *ceteris paribus* condition, the same conditionals in counterfactual form are not. That is, Pearl thinks of conditional sentences as enthymemes[107]. When such sentences are evaluated in the context of a causal model many of the missing details that are intended to attach to the antecedents of such (predictive) conditionals, but that remain obscure in the standard presentation, are made explicit. For Pearl the counterfactuals consistent with a causal model therefore are conditional sentences whose logical form makes explicit what would otherwise remain enthymematic. The important point to note is that the nature of counterfactuals in Pearl's account is tied to the evaluation of predictions under (simulated) experimental conditions and, thus, that Pearl assimilates counterfactuals to a species of indicative conditionals. On reflection of this fact Pearl may have been better to devise an alternative label for the generic counterfactual form that plays a central role in his semantics for causal models. Calling such conditionals 'predictive hypotheticals' would be more in keeping with the spirit of the account especially in light of the fact that the truth-conditions of such sentences are relativised to models and not to either the utterances of scientists or the truth-conditions of such utterances.

As is the case with Pearl's conception of counterfactuals, Pearl's conception of type and token causes is somewhat unconventional. It is important to discuss Pearl's view of types and tokens in order to provide some specification of the information expressed by a causal model.

As I note above, one important component of causal models is the representation of mechanisms by functions. According to Pearl these functions supply information at both the type and the token level:

> These functions are type-level in the sense of representing generic, counterfactual relationships among variables that are applicable to every

[106] And, hence, cannot straightforwardly be seen to pertain to concrete facts since the relation between models and reality is as yet undefined.

> hypothetical scenario, not just the ones that were realised. At the same time, any specific instantiation of those relationships represents a token-level claim (Pearl 2000a: 310).

That is, any given set of functions can be taken to represent a type of causal scenario–one where the variables are related according to the specification of the set of functions–and the specification of that scenario with actual values of each of the variables would represent a token of that scenario type. What differentiates one scenario from another is the level of detail available to the specification of the background variables U. A full specification of U turns a causal model into a causal world and allows causal claims to be assessed at the token level. However, for Pearl it is more often the case that the investigator does not

> ...possess the detailed knowledge necessary for specifying a single world $U = u$, and [uses] a probability $P(u)$ to summarise [their] ignorance of those details. This takes [the investigator] to the level of probabilistic causal models $\langle M, P(u) \rangle$. Causal claims made on the basis of such models, with no reference to the actual scenario, would be classified as type-level claims (Pearl 2000a: 310).

But, often the investigator will possess some information relevant to the scenario under study. Pearl (2000a) calls this information *evidence* and asserts that it may be used to update $P(u)$ into $P(u \mid e)$ such that causal claims derived from the model $\langle M, P(u \mid e) \rangle$ represent token claims of varying degrees, depending on the specificity of the evidence. Hence, on Pearl's account the difference between type and token causal claims is a matter of degrees such that the higher the level of specific evidence available to the investigator the closer the investigator comes to making token claims and hence statements about *actual* causes (Pearl 2000a: 311). Pearl defines an actual cause in the following way:

---

[107] For a discussion of enthymemes relevant to the present issues see Priest (2001: 77).

Actual Cause:

$\vec{X} = \vec{x}$ is an actual cause of φ in $(M,\vec{u})$ if the following three conditions hold:

(AC1) $(M,\vec{u})$ semantically entails $(\vec{X} = \vec{x}) \wedge \varphi$. That is, both $\vec{X} = \vec{x}$ and φ are true in the actual world.

(AC2) There exists a partition $(\vec{Z}, \vec{W})$ of V with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}',\vec{w}')$ of the variables in $(\vec{X},\vec{W})$ such that if $(M,\vec{u})$ semantically entails $Z = z^*$ for $Z \in \vec{Z}$ then

(a) $(M,\vec{u})$ semantically entails $\left[\vec{X} \leftarrow \vec{x}',\vec{W} \leftarrow \vec{w}'\right]\neg\varphi$. In words, changing $(\vec{X},\vec{W})$ from $(\vec{x},\vec{w})$ to $(\vec{x}',\vec{w}')$ changes φ from true to false,

(b) $(M,\vec{u})$ semantically entails $\left[\vec{X} \leftarrow \vec{x},\vec{W} \leftarrow \vec{w}',\vec{Z}' \leftarrow \vec{z}^*\right]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$. In words, setting $\vec{W}$ to $\vec{w}'$ should have no effect on φ as long as $\vec{X}$ is kept at its current value $\vec{x}$, even if all the variables in an arbitrary set of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

(AC3) $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions (AC1) and (AC2). Minimality ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing φ in (AC2)(a) are considered part of the cause; inessential elements are pruned Halpern and Pearl 2001a).

Where, given a signature $S = (U, V, R)$[108], a formula of the form $X = x$, for $X \in V$ and $x \in R(X)$, is called a *primitive event*. And a *basic causal formula* (over S) is one of the form $[Y_1 \leftarrow y_1,\ldots,Y_k \leftarrow y_k]\varphi$ where φ is a Boolean combination of primitive events, $Y_1,\ldots,y_k$, X are variables in V, with $Y_1,\ldots,y_k$ are distinct, $x \in R(X)$, and

[108] A signature S is a tuple (U, V, R), where U is a finite set of exogenous variables, V is a finite set of endogenous variables, and R associates with every variable $Y \in U \cup V$ a nonempty set $R(Y)$ of possible values for Y (the set of values over which Y ranges). A causal model over a signature then is a tuple $M = (S, F)$, where F associates with each variable $X \in V$ a function denoted F_X such that F_X tells us the value of X given the values of all the other variables in $U \cup V$.

$y_i \in R(Y_i)$. And where such a formula is abbreviated as $\left[\vec{Y} \leftarrow \vec{y}\right]\varphi$. A *causal formula* is a Boolean combination of basic causal formulas[109]. Given this syntax, a causal formula ψ is true or false in a causal model given a *context*. Pearl writes $(M,\vec{u})$ semantically entails ψ if ψ is true in a causal model M given a context $\vec{u}$[110]. $(M,\vec{u})$ semantically entails $\left[\vec{Y} \leftarrow \vec{y}\right](X = x)$ if the variable X has the value x in the solution to the equations in $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$. And $(M,\vec{u})$ semantically entails $\left[\vec{Y} \leftarrow \vec{y}\right]\varphi$ for any arbitrary Boolean combination φ of formulas of the form $\vec{X} = \vec{x}$ is defined similarly Halpern and Pearl 2001a). The general idea, as Hopkins and Pearl (2003) put it, is that x is an actual cause of y just when x and y are the actual values of X and Y (AC1) and under a specific counterfactual contingency w, the value of Y is dependent on X in such a way that setting X to its actual value ensures that Y remains at its actual value even in the case that all other variables in the model are fixed at their actual values (AC2), so long as the set of actual values of X have no values that fail to sustain the effect (AC3)[111].

The following version of the desert traveller scenario illustrates well the move from general to actual causal claims. Consider the following model constructed by an investigator who has undertaken to uncover what caused the death of a desert traveller. All the investigator knows of the circumstances surrounding the travellers death is that two assassins were dispatched, each of whom claim to have caused the travellers death; one by shooting a hole in the travellers canteen and the other by poisoning the travellers canteen with cyanide. The investigator does not know the results of an autopsy later carried out on the traveller nor whether the traveller managed to drink from the canteen before it was emptied. The investigator, reasoning that the traveller died from either dehydration or cyanide poisoning conceives the following model consisting of six variables; y = desert traveller's death, D =

---

[109] For further discussion and interpretation of actual causality see Halpern and Pearl (2001a) and chapter 10 of Pearl (2000a).

[110] That is, a context is a specific setting of variables in U to some value $\vec{u}$.

[111] For further discussion of Pearl's definition of actual causality, especially condition (AC2)(b) see Menzies (2002), Halpern and Pearl (2001a), Hopkins and Pearl (2003).

dehydration, C = cyanide intake, X = assassin 1's attempt to shoot the traveller's canteen, P = assassin 2's attempt to poison the traveller's canteen, u = the time elapsed before the traveller's first drink from the canteen[112]. The value of y is known to be true as is the value of X and P respectively. The value of u is not known by the investigator and so has to be estimated. For Pearl this amounts to summarising the investigator's ignorance of u by $P(u)$. The model has the following functions:

$$c = p(u' \vee x')$$
$$d = x(u \vee p')$$
$$y = c \vee d$$

The investigator must then calculate the respective probabilities that either one of c or d was the cause of the traveller's death by elaborating on the effect that different values of u have on the function of $y = c \vee d$. The state $u = 1$ denotes the event that the traveller did not reach for a drink before assassin 1 holed the traveller's canteen. Given this setting the investigator notes the model's functions as:

$$c = x',$$
$$d = x,$$
$$y = d,$$

and proceeds to test whether x or p was the cause of y. In the instance the model identifies assassin 1's actions as the actual cause of the traveller's death. It is easy to see that in the event that the traveller had a drink before assassin 1 had time to shoot the canteen (i.e. $u = 0$), the model identifies the actions of assassin 2 as the actual cause of the traveller's death. As Pearl notes, without knowledge of which state actually prevailed, the investigator must settle for the probability that x caused y. The example illustrates how the investigator is able to reason about the general tendency of x to cause y in the given scenario and that the move towards the identification of the actual cause of y is limited by the total evidence available (Pearl 2000a: 310).

[112] Plus two distinct variables that summarise the causes behind each assassin's actions.

In ending the outline of Pearl's account I admit to neglecting a number of details. Some of the details I consider important but that remain outstanding include: the nature of the relation between mechanisms and laws; the relation between models and reality; the distinction between statistical and causal analysis; and the role that pragmatic issues play in model construction. There are two main reasons for not addressing such details here. First, each of these issues are, I think, far from clear as they are presented in Pearl (2000a) and as such require a good deal of space to adequately address and, second, each of these issues deserve a lengthy independent treatment. Unfortunately, to do justice to each is well beyond the present forum. However, one issue cannot go by without investigation. The issue involves the distinction Pearl (2000a) draws between various statistical and causal notions. I deal with this in the following section.

2.1 The Poverty of Statistics

In this section I present and discuss a collection of Pearl's views on the boundary between statistical modelling and causal modelling. I set these views out more or less as they are found in Pearl (2000a) except for the fact that here they are collected together whereas there they are scattered throughout the book. The reason for including this section is to add further crucial detail beyond that which I have already afforded to Pearl's theory of causality before I offer my interpretation of the theory in the following chapter. Important detail would be missing were an interpretation to proceed without taking note of Pearl's views on the boundary between 'the statistical' and 'the causal.'

The first issue to discuss involves Pearl's views on the nature of probability and his commitments to Bayesian statistics. Pearl (2000a) presents his account from a Bayesian perspective and so treats probabilities as degrees of rational belief expressible as propositions[113]. For the sake of simplicity Pearl assumes there is no relevant difference between sentential propositions and the 'events' denoted by propositions. Consequently Pearl has no truck with evaluating probabilistic statements as either true or false. Pearl's commitment to a subjective Bayesian interpretation of probability extends also toward Bayesian statistical methods.

As I have mentioned above, the basic expressions in the Bayesian canon are statements about conditional probabilities. Typically, the specification of a conditional probability statement is interpreted as the degree of belief in some event A under the assumption that another distinct event B is known with certainty. Bayesians think that this value may in turn provide guidance in assigning degrees of belief to probability statements that involve joint events, or that may be independent, or that may be conditionally independent. In short, conditional probability statements

[113] Pearl's account is primarily set out for systems comprising a finite number of discrete variables. The account may be extended to handle continuous variables with the addition of the appropriate

are taken by Bayesians to be more fundamental than probability statements concerning joint events, since, Bayesians argue, conditional relationships are more compatible with the organization of human knowledge (Pearl 2000a: 3).

A related characteristic of Bayesian methods concerns the status of Bayes' inversion formula:

$P(H \mid e) = P(e \mid H)\, P(H) / P(e)$.

Bayesians take the inversion formula to state that the belief accorded to a hypothesis H upon obtaining evidence (e) is the result of multiplying one's previous belief or prior probability P(H) by the likelihood or posterior probability P(e | H) that such evidence will appear if the hypothesis H is true. As Pearl states;

> The Bayesian subjectivist regards the inversion formula as a normative rule for updating beliefs in response to evidence. In other words, although conditional probabilities can be viewed as purely mathematical constructs, the Bayes adherent views them as primitives of the language and as faithful translations of the English expression '..., given that I know A' (Pearl 2000a: 5-6).

However, Pearl (2001a) is at pains to qualify his commitment to Bayesian principles. Pearl holds the following three assertions to express several core principles of Bayesian statistics:

> (i) When reasoning about the world one cannot do so in ignorance of one's own knowledge of the world.
> (ii) It is natural and useful to cast what we know about the world in the language of probabilities.

---

mathematical machinery. Pearl adheres to the Kolmogorov axiomatization of the probability calculus (Pearl 2000a: 2-3).

(iii) If our subjective probabilities are erroneous, their impact will get washed-out in due time as the number of observations increases[114].

As is clear from the prior points Pearl accepts (i). But Pearl expresses reservations about (ii) due principally to the fact that Pearl thinks there is a point at which reasoning about the world with probabilities and evidence alone becomes unproductive. Pearl doubts (ii) because Pearl believes causality is different (though not unrelated) in formality and subject matter to probability and that causality is the more fundamental notion of the two. I discuss this further directly. In turn, the reasons behind Pearl's reservations about (ii) hold negative consequences for the prospect of accepting (iii). Pearl holds that (iii) is false. One reason given by Pearl is that one cannot expect prior probabilities and evidence about a system to lead to correct estimates and sound inferences of and from the systems parameters unless one has identified the 'correct' measures. Pearl does not think that correct measures are forthcoming without resort to non-statistical (i.e. causal) information. In other words, Pearl denies that techniques based on probability theory alone can guarantee the reliability of estimates and inferences about systems involving causal relationships.

Pearl's reservations about Bayesian statistical methods hinge on their being a meaningful difference between causal and statistical notions. Indeed, at several points throughout his (2000a), Pearl is at pains to establish a boundary between causal concepts, and statistical concepts. Pearl dedicates several discussions to the topic of the 'fundamental' nature of causal parameters in contrast to the 'superficial' nature of statistical and probabilistic parameters. The primary reason behind Pearl's belief that there is a boundary between the two is that concepts we commonly and un-controversially take to be causal in nature cannot be expressed in terms of probabilities alone. One example Pearl often cites to illustrate this point involves the statistical dependence of mud on rainfall. Pearl asserts that one may assess the relationship between mud and rainfall using only statistical methods and expressions

[114] For a deeper discussion of Bayesianism see De Finnetti (1990), Howson and Urbach (1993), Kyburg and Smokler (1980).

so long as no attempt is made to provide the relationship with a causal sense. Pearl's point is that there is no way of using the probability calculus to express the fact that mud does not cause rain, and, because of this, no other causal relationships between rainfall and mud can be read into the statistical dependence of one upon the other. But this is despite the fact that it is natural to interpret the dependence of mud on rainfall in a (pre-theoretic) causal sense. Pearl seeks to generalise this point across the sorts of associations commonly displayed throughout economics and the social and biological sciences[115].

In order to illustrate this point in more detail it is necessary to first introduce Pearl's definitions of probabilistic, statistical (associational) and causal parameters, and statistical and causal assumptions and concepts respectively.

Pearl thinks of a probabilistic parameter as any quantity defined in terms of a joint probability function. A statistical parameter is, in turn, any quantity defined in terms of a joint probability distribution of observed variables and where no assumptions are made about the existence of unobserved variables. Pearl thinks of a causal parameter as any quantity defined in terms of a causal model that is not already a statistical parameter. Speaking to these definitions Pearl remarks that:

> The distinction between probabilistic and statistical parameters is devised to exclude the construction of joint distributions that invoke hypothetical variables (eg counterfactual). Such constructions, if permitted, would qualify any quantity as statistical and would obscure the distinction between causal and non-causal assumptions (Pearl 2000a: 39).

Furthermore, Pearl thinks a *statistical assumption* is any constraint on a joint distribution of observed variables and a *causal assumption* is any constraint on a causal model that cannot be realised through the imposition of statistical assumptions.

---

[115] For an analogous case made in terms of the deliberation on possible actions see Pearl (2000a: 108-109).

Even so, Pearl admits it is possible that causal assumptions may have statistical implications, in which case Pearl calls the causal assumptions 'testable' or 'falsifiable' (Pearl 2000a: 38-39).

Pearl further elaborates the boundary between the causal and the statistical by providing examples of concepts, familiar from the field of statistics generally, which he thinks fall into one category but not the other. According to Pearl the following concepts are statistical: correlation, regression, conditional independence, association, likelihood, collapsibility, risk ratio, odds ratio. The remaining concepts are for Pearl causal: randomisation, influence, effect, confounding, exogeneity, ignorability, disturbance, spurious correlation, path coefficients, instrumental variables, intervention and explanation (Pearl 2000a: 40). The reader may substitute the word 'associational' for 'statistical' without loss of meaning.

Although Pearl claims the distinction between the probabilistic and statistical on the one hand and the causal on the other is clear and distinct he nevertheless thinks the two remain closely allied and, in fact, that the latter is an extension of the former[116]:

> Thus, I have tried in [setting out my theory] to present mathematical tools that handle causal relationships side by side with probabilistic relationships (Pearl 2000a: xiv).

And:

> The purpose of this demarcation line is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators to treat non-statistical concepts with the proper set of tools (Pearl 2000a: 40).

In light of this demarcation between the statistical and the causal Pearl goes on to make several claims about the epistemology and ontology of causality. Pearl's views

appear to represent a reversal of at least one key characteristic of the statistics paradigm in that, on Pearl's account, it is causal relationships that are assumed to be the fundamental constituents of the world and which generate probabilistic relationships, not vice versa. Call this, for present purposes, 'Pearl's reversal.' The position Pearl adopts towards the border between statistics and causality can be summarised in the following way: Pearl thinks of the world as containing many actual physical processes that span the range from deterministic to indeterministic depending on the specific detail with which such processes are taken into account. Many of these processes are either unobservable or unmeasurable. The collection of these processes and their properties stand (for the most part) independent of human tastes and interests. However, the world is not closed to investigation since it may be manipulated and observed. Humans are capable of direct causal interaction with parts of the world and may possess in specific instances causal knowledge that is not itself merely the convenient linguistic abbreviation of perceived associations. Humans are further capable of learning and storing causal information either as the result of direct interaction with the world or by being taught by others. Typically this information is stored in the form of (some number of related) counterfactual propositions. Batteries of such counterfactuals will be true just in case the causal model, of which they form a part, offers an adequate representation of the behaviour of the physical circumstances it models.

Hence, for Pearl the main difference between statistical modelling and causal modelling is the belief that the latter begins where the former ends (despite some commonalty of formalism). For Pearl (2000a), this means that statistical modelling is limited to the estimation and manipulation of expressions that represent what Pearl thinks of as 'static' observations, whereas causal modelling contains the facility to extend manipulation to the mechanisms responsible for the production of those observations. Clearly the difference spans both the formal and subject matter

[116] See also the discussion of how probability relates to causality on pp 1-2 of Pearl (2000a).

components of statistical and causal modelling respectively but seems driven primarily by the latter[117].

A consequence of Pearl's attempt to draw a line between the statistical and the causal is that he rejects attempts to reduce causal relationships to probabilistic relationships. This holds negative implications for the project of providing a probabilistic account of causality. The nature of the implication necessitates a closer look at exactly what Pearl envisages the relationship between the subject matter of causality and probability to be.

Pearl considers the project of providing a probabilistic account of causality to be characterised by the attempt to explicate causality solely in terms of probabilistic relationships. When discussing the possibility of providing a probabilistic account of causality Pearl (2000a) does not draw any explicit distinction between probabilistic theories of causality and theories of probabilistic causality. Even so, it is relatively clear Pearl intends to refer to probabilistic theories of causality since such theories: (i) claim that causes are probability raisers and, hence; (ii) causal concepts are reducible to probabilistic concepts[118]. Indeed, in his discussion of probabilistic theories of causality Pearl (2000a) refers primarily to the work of Cartwright and Eells[119]. Pearl's main criticism of this project is that it cannot be completed without circularity. Hence, the best that may be achieved by entertaining a probabilistic theory of causality is, according to Pearl, the construction of a consistency testing procedure for comparing sets of causal statements with available (temporal and) probabilistic information (Pearl 2000a: 251). On this point Pearl remarks:

[117] Elsewhere on this point, Pearl adds: "Even the theory of stochastic processes, which provides probabilistic characterisations of certain dynamic phenomena, assumes a fixed density function over time-indexed variables. There is nothing in such a function to tell us how it would be altered if external conditions were to change. If a parametric family of distributions is used, we can represent some changes by selecting a different set of parameters. But we are still unable to represent changes that do not correspond to parameter selection; for example, restricting a variable to a certain value, or forcing one variable to equal another" Pearl (2001: 28 n1).
[118] For instance, see comments on p 249 of Pearl (2000a).

> ... the basic program of defining causality in terms of conditionalisation, even if it turns out to be successful, is at odds with the natural conception of causation as an oracle for interventions. This program first confounds the causal relation $P(E \mid do(C))$ with epistemic conditionalisation $P(E \mid C)$ and then removes spurious correlations through steps of remedial conditionalisation, yielding $P(E \mid C,F)$. The structural account, in contrast, defines causation directly in terms of Nature's invariants (Pearl 2000a: 252)[120].

Humphreys draws the connection between the notion of statistical correlation and causation:

> There is no purer case of an empiricist association between events than a statistical correlation, and when such associations [...] concern occurrent events, they are the obvious analogues of classical regularity accounts of causation (Humphreys 1989: 50).

It is just this commonality that Pearl has in mind when he lumps statistical assumptions together with the project of probabilistic causality under the same criticism, which has been dubbed 'passive empiricism' (Humphreys 1989: 47). According to passive empiricism, humans are a device capable of collecting and assessing observations so as to identify particular types of regularities. Humphreys (1989) provides the following summary of the passive empiricist's position:

> The essence of passive empiricism is the view that humans are special kinds of receiving devices capable of assessing observed data for empirical properties, that in the case of Humean causation, for example, would be regular temporal succession, spatiotemporal contiguity, and the kind of regular association that constitutes a lawlike regularity. Within Hume's own

[119] In particular Cartwright (1979) and Eells (1991). Pearl also mentions the work of Spirtes in this context.

> account of sensory impressions, for example, what produces the impression is irrelevant. [...] This disregard for the origins of the impressions is reasonable within a certain kind of empiricism. If one thing that motivates you to be an empiricist is the desire to remain epistemically conservative, then to avoid moving beyond the security of the immediately given to inferred entities is desirable, because such inferences are fallible; alternatively, if such inferences require a move from an observed effect (the impression) to an unobserved cause (the source), one might object to the circularity of such a move within the context of an empiricist analysis of causation. But whatever the reasons, the analysis is conducted at the level of the observations, and the source of the empirical data is not a part of the analysis itself (Humphreys 1989: 47).

Pearl (2000a) clearly agrees with Humphreys' last assertion. Pearl criticises passive empiricism for what he calls the 'closed world assumption.' Pearl claims the closed-world assumption is "the most critical and least defensible paradigm underlying probabilistic causality" (Pearl 2000a: 252). According to Pearl:

> [...] probabilistic causality rests on the assumption that one is in the possession of a probability function on all variables in a given domain. This assumption absolves the analyst from worrying about unmeasured spurious causes that (physically) affect several variables in the analysis and still remain obscure to the analyst. [...] Because they are unmeasured (or even unsuspected), the confounding factors in such examples cannot be neutralised by conditioning or by 'holding them fixed.' Thus, taking seriously Hume's program of extracting causal information from raw data entails coping with the problem that the validity of any such information is predicated on the

[120] Where E is thought to be the effect, C the cause and F is a particular truth value assignment to variables in some background context K supposedly not containing C or E. By structural account read Pearl's account of causality.

> untestable assumption that all relevant factors have been accounted for (Pearl 2000a: 252)[121].

Pearl continues on from this point to argue that human beings have a capacity for manipulative experimentation of their environment as well as the ability to communicate causal information that the passive empiricist is denied due to the assumptions (such as those just outlined) that underlie their position. This sets up what is probably the defining feature of Pearl's disagreement with empiricism[122]; Pearl holds probability to be epistemic, but thinks of causation as an objective feature of the world.[123]:

> I now take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world (Pearl 2000a: xiii-xiv).

And,

> [...] causal relationships are ontological, describing objective physical constraints in our world[...] (Pearl 2000a: 25)[124].

---

[121] I am not certain of what Pearl means by the assertion that Hume held a program to extract causal information from observations since it is well known that Hume attempted to show that just such a program could not succeed.

[122] However, note that even though Pearl disagrees with empiricism he found it necessary to argue that counterfactuals have empirical content.

[123] However, Pearl does countenance a theory of objective chance but it is unclear whether he holds that probabilities are truly objective in this instance or merely offer a model of chance. See Pearl (2000a: 220-221) for discussion. I do not wish to beg any questions here against empiricism. Many empiricists believe causation to be an objective feature also, so long as one means by that nothing more than regularity. The salient point is that Pearl appears to be asserting that causal relations are something more than mere regularity.

[124] Clearly there is more going on here than meets the eye. Ultimately, identifying Pearl's stance on such issues will depend upon whether or not he endorses the Humean Supervenience Thesis. I note that this is a different issue from ascertaining whether or not Pearl's account requires the Humean Supervenience Thesis. I return to this issue in chapter 3.

Moreover, as I described in section 1.3.3 Pearl holds that the functions of a causal model may be read as law-like mechanisms and causal models offer a partial or complete description of reality:

> The basic building blocks of the structural account are the functions, which represent lawlike mechanisms [and the] ingredients that distinguish one [causal] scenario from another are represented in the background variables U. When all such factors are known, $U = u$, we have a 'world' on our hands–an ideal, full description of a specific scenario in which all relevant details are spelled out and nothing is left to chance or guessing. Causal claims made at the world level would be extreme cases of token causal claims (Pearl 2000a: 310)[125].

With these points in mind it is necessary to reflect momentarily on the relationship between the statistical (and thus probabilistic) component of Pearl's account and its causal components. The thought occurs that the distinction Pearl has drawn between the causal and the statistical is so strong that his own account of causality and causal discovery may fall foul of it.

It is fair to say the primary reason Pearl draws a distinction between statistical modelling and causal modelling is that statistical modelling is limited to working with observations, whereas causal modelling reaches deeper to work with the mechanisms that produce observations. This view apparently confers causal models with an objective character that statistical models lack. The reason Pearl thinks of the causal as in some way more objective and fundamental than the statistical is, I think, due to Pearl's conception of manipulative experimentation.

> Manipulation subjugates the putative causal event to the sole influence of a known mechanism, thus overruling the influence of uncontrolled factors that

> might also produce the putative effect. [...] The whimsical nature of free manipulation replaces the statistical notion of randomised experimentation and serves to filter [observations] produced by [interventions] from those produced by uncontrolled environmental factors (Pearl 2000a: 253).
>
> This understanding of causal influence permits us to see precisely why, and in what way, causal relationships are more 'stable' than probabilistic relationships. [...] Causal relationships should remain unaltered as long as no change takes place in the environment, even when our knowledge about the environment undergoes changes (Pearl 2000a: 25).

Pearl thinks that manipulation offers a method for revealing and correctly describing natural mechanisms (Pearl 2000a: 25, 253) and it is only 'stable' probabilistic relationships that are the manifestation of natural mechanisms (Pearl 2000a: 63). Pearl denies that the facility exists for the description of such relationships within the statistical paradigm. Shipley (2000) captures this world-view with the following analogy. Consider a puppet show where a puppet master tells stories by manipulating three-dimensional puppets behind a screen and in front of a bright light. The puppets intercept the light and cast two-dimensional shadows onto the screen for the audience to see and interpret. In order to infer the three-dimensional action the shadows must be detailed and placed in context. Shipley likens the puppet show to the scientists attempt to infer causality from patterns of association:

> Biologists are unwitting participants in nature's shadow play. These shadows are cast when the causal processes in nature are intercepted by our measurements. Like the audience at the [puppet show], the biologist cannot simply peak behind the screen and directly observe the actual causal processes. All that can be directly observed are the consequences of these processes in the form of complicated patterns of association and independence

[125] Furthermore, Pearl holds that humans possess some form of innate inferential machinery that enables the storage and processing of causal information in the form of counterfactuals. For instance

> in the data. As with shadows, these correlational patterns are incomplete–and potentially ambiguous–projections of the original causal processes. As with shadows, we can infer much about the underlying causal processes if we can learn to study their details, sharpen their contours, and especially if we can study them in context" (Shipley 2000: 1).

Hence, Pearl (2000a) accepts that correlation often implies causation and that causation fixes correlations, that is, causal relationships between objects or variables determine the correlational relationships between them (Shipley 2000: 1-2; Pearl 2000a: 59-60). These views, being fundamental to Pearl's account, lead it to a difficulty I shall now discuss.

Recall that Pearl interprets probabilities as degrees of belief and explains that Bayesian Networks gain their name (in part) due to the subjective nature of their input information (Pearl 2000a: 2, 14). Further recall that Pearl demonstrates the conditional dependencies of a distribution and the functions of a causal model representing that distribution are equivalent. That is, Pearl has shown that for every Bayesian network G and distribution P, there is a functional model that generates a distribution identical to P, and hence that functional models are another way of encoding joint distribution functions (Pearl 2000a: 30-31). These facts tie the statistical portion of Pearl's account very closely to the causal portion of Pearl's account. In particular, the subject matter of the probabilities in a Bayesian network is carried over to causal models. This is *prima facie* puzzling given the fundamental differences Pearl thinks exist between the statistical and the causal. In a nutshell, it is a puzzle how Pearl is able to move from a statistical interpretation of the relationships within a Bayesian Network to a causal interpretation of the same network when the former encodes degrees of belief and the latter encodes mechanisms, which represent constraints on objective law-like relations in the natural world. Glymour (2003) presents an argument which I will exploit to show that Pearl cannot both solve this

---

see Pearl (2000a: 253, 223-224). I return to this issue in chapter 3.

puzzle and maintain a subjective interpretation of probability without giving up the view that causal models are about objective constraints in the world.

Consider that probability is combined with science in two distinct ways. One is when probabilities become part of the content of science such as in statistical mechanics and quantum mechanics. The other is when probability is utilised by science to justify methods of inference from observation. Define instrumentalism about some part of language as the doctrine that there are sentences within it that are not claims about the way the world is or could be but linguistic devices for making inferences about other sentences that are claims about the way the world is or could be. On this definition, a probability claim is instrumental if it says nothing about what happens or what could happen in the domain under scientific study but plays a role in licensing inferences that do say something about what happens or what could happen (Glymour 2003: 237, 240).

Consider the claims of several common interpretations of probability. The probability claims of the subjective Bayesian, limiting frequentist and propensity theorist say nothing about what happens or what could happen. For instance, taking each interpretation in turn, the subjective Bayesian says nothing about what happens or could happen because the Bayesian's subject matter is norms of belief as distinct from, say, data points. For the limiting frequentist, probability claims are statements about limiting relative frequencies of a property in an infinite sequence. However, since the latter are consistent with any claim about any finite set of events whatsoever, the limiting frequency interpretation of probability fails to say anything about what happens or what could happen. Propensity theorists propose an unmeasurable physical property that neither precludes nor necessitates any occurrence. Hence, the propensity interpretation of probability does not entail any claim about what happens or what could happen (Glymour 2003: 238).

To get probability claims–whether subjective, limiting frequentist, or propensity–to support inferences from data, the accepted method is to follow a prescribed route that

first takes the investigator through a process of estimation and then through a process of decision. The former moves from measures of nature to probability and the latter moves from probability to judgements about nature (Glymour 2003: 240-241). But, as Glymour points out, stepping this route displaces the general goal of discovering what is happening in the world:

> Mathematical statistics and decision theory [...] clarify immensely the role of probability in the analysis of data. There is a price. The justification for the [route taken through procedures for estimation and decision making] is that given the evidence, given the supposition that the reasoner was rational before acquiring the evidence, and given the utilities of the reasoner, the judgements that result are required by rationality. The clarification requires a change in the primary goal of inquiry from the pursuit of truth to the pursuit of rationality (Glymour 2003: 240-241).

Now recall that Pearl draws a connection between causality and what he labels variously 'objective constraints' and 'laws'. The claim Pearl appears to make here is that successfully estimating causality is in part to uncover the objective constraints that govern the specific system under investigation. This seems to relate the notion of intervention with that of invariance[126]. As Woodward explains:

> [...] one may think of an intervention as an idealisation of an experimental manipulation carried out on some variable *X* for the purpose of ascertaining whether changes in *X* are causally or nomologically related to changes in some other variable *Y* (Woodward 2000: 199).

> [...] a generalisation describing a relationship between two or more variables is invariant if it would continue to hold–would remain stable or unchanged–as various other conditions change (Woodward 2000: 205).

In other words, Pearl's theory is designed for estimating causal quantities in manipulated and un-manipulated systems, or put simply, Pearl's theory is designed to say something about the way the world is, or could be. But, since Pearl attempts to do so using one of the three offending interpretations of probability mentioned above, Pearl (2000a) becomes a target for Glymour's argument. First, because Bayesian Networks encode degrees of belief Bayesian Networks do not say anything about what happens or what could happen. At best, Bayesian Networks speak about what the investigator is rationally compelled to accept given the evidence. Since causal models are equivalent to Bayesian Networks (in the salient sense), causal models do not say anything about the way the world is or could be. It follows immediately that causal models do not say anything about objective physical relationships in nature. To paraphrase Glymour (2003), in adhering to a subjective interpretation of probability, Pearl sets aside the primary goal of arriving at the true values of causal quantities in favour of the rational acceptability of estimates. But this conclusion, of course, conflicts with Pearl's interpretation of causal models. The puzzle remains; how can the causal models of Pearl's theory be about physical mechanisms when the probability expressions of such models are about norms of belief? Pearl's causal models do not appear to be consistent with the distinction Pearl enforces between the statistical and the causal.

It is my view that this difficulty is rooted in Pearl's subjective interpretation of probability. However, one may expect Pearl to resolve the puzzle by showing it is possible to maintain both subjective and objective components of causal models via some appropriate form of bridging principle. For instance, Pearl might choose to conform his subjective probabilities to a principle of direct probability or to something like Miller's Principle[127]. Pearl might also seek to maintain a dualist position about the application of probability such as that expressed by Levi (2003).

---

[126] I discuss these notions more thoroughly in chapter 3.

[127] Although Pearl does not discuss such bridging principles directly Pearl's comments regarding human psychological capacities for causal identification and inference are suggestive. I discuss such issues in the following chapter. For discussion of direct probability and related issues see Hajek (2003), Hacking (1965), Kyburg (1974), Levi (1977), Williams (1963), Horwich (1982), and Levi (2003). For discussion of Miller's principle see Lewis (1980).

However, I suggest the solution offered by Glymour (2003) is at once implicit in and in agreement with the intended application of Pearl's theory (minus its subjective component)[128].

Glymour argues that the theory of probability provides instruments for making valid inferences about the values of quantities and the truth and falsity of hypotheses rather than valid inferences about which decision has the highest expected utility (Glymour 2003: 251). The account of probability Glymour thinks is instrumental in drawing such inferences is the long-run frequency interpretation, which says the probability of an outcome in a trial is approximately the relative frequency of the outcome in a 'long' (but not infinite) sequence of trials. Glymour suggests that probabilities on the finite frequency interpretation are appropriate instruments for drawing inferences about what happens or what could happen in the world so long as we understand the interpretation

> [...] as a proposal to use the language and mathematics of probability to approximately describe actual or potential finite populations, and as a means of generating definite, nonprobabilistic hypotheses (Glymour 2003: 249).

Glymour argues that the benefit of adopting the finite frequency interpretation is that the degree of approximation in an approximate finite frequency claim can, unlike the probability claim itself, be made explicit and empirical (Glymour 2003: 249). According to Glymour, once made explicit,

> [...] 'probability' claims become assertions about bounds on arrangements of values of quantities in finite populations. They become an especially interesting variety of claims, not about probabilities, but about uncertain but bounded errors in finite frequency distributions. Entirely explicit versions of finite frequency claims of probability, on my analysis, are claims about the

[128] Especially so since Pearl claims to afford practitioners a method of reasoning about causal queries that contain no probabilistic elements.

> uncertain but bounded error of some function of the empirical distribution of a quantity (or quantities) in an actual or potential finite population (Glymour 2003: 249-250).

In terms of Shipley's shadow analogy, one may interpret the shadows to represent finite frequency distributions (or a function of the distribution of a quantity in an actual or potential finite population), and the audience's inference to represent definite hypotheses. In the context of a causal model such hypotheses are (partly) encoded by the missing arrows of its DAG and are about what form finite frequency distributions take under manipulation.

It remains to ask what becomes of the subjective and the objective characteristic of causal models. I suggest, in light of the forgoing discussion, two distinct senses of causality are employed by Pearl (2000a). One of the two senses is discernable from Pearl's comments regarding the psychology of causal knowledge and from Pearl's views on the transition from Bayesian Networks to causal models. A distinct sense of causality is discernable from Pearl's comment that mechanisms are to be interpreted as objective constraints on physical processes. Consider, for instance, the following comments:

> A causal model is not just another scheme for encoding probability distribution through a set of parameters. When we come to define mathematical objects such as causal models, we must ensure that the definition captures the distinct ways in which these objects are being used and conceptualised (Pearl 2000a: 63).

> Our objective is to preserve, explicate, and satisfy–not destroy–[scientist's causal] intuitions (Pearl 2000a: 26).

> The often heard argument that human intuitions belong in psychology and not in science or philosophy is inapplicable when it comes to causal intuition –

> the original authors of our causal thoughts cannot be ignored when the meaning of the concept is in question. Indeed, compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation, and the proper incorporation of background information into statistical studies likewise relies on accurate interpretation of causal judgement (Pearl 2000a: 26 n12).
>
> [The] mechanism-based conception of interventions provides a semantical basis for notions such as 'causal effects' or 'causal influence,' [...] (Pearl 2000a: 24).
>
> [...] it is no wonder that people prefer to encode knowledge in causal rather than probabilistic structures (Pearl 2000a: 25).

It is, I think, natural to read such comments to refer to the *concept* of causality, especially as the concept is countenanced within the sciences[129]. I see these comments as an explicit statement of the underlying principles involved in the use or employment of causal concepts (Hart and Honore 1985: 26). That is, one might appropriately make these sorts of comments in case one wishes to (partly) explain the *meaning* of the term 'causality' in the context of modelling in science. Likewise for the step Pearl takes from an associational reading of graphs to the causal reading. This step is taken on the assumption that one interprets the functions of graphs to represent the sorts of apparently natural relations (Pearl believes are) connoted by the term 'causality'. The remaining sense is then to be identified with Pearl's assertions that causality is an objective feature of the world and that the mechanisms of his account are representatives for law-like relations in nature.

A distinction I cited in chapter 1 helps to clarify the ambiguity. I said there that theories of causation may be ordered according to one of two categories. In one

[129] See also comments made in the same spirit throughout chapter 10 of Pearl (2000a) especially pp 325, 327, 328.

category belong those theories that analyse causality against human concepts and practices. In the other category are those that analyse causation without (any intended) recourse to human interests or tastes and instead aim to describe what causality is in nature. Put simply, those in the former category seek to elucidate what people mean by causal terms whilst those in the latter seek to analyse what causation is in the objective world. It strikes me that the two senses of causation I identify in Pearl's account coincide with this distinction. I investigate the possibility that Pearl offers a two-part theory of causality in the following chapter.

3.0 Introduction

My aim in this chapter is to provide a philosophical interpretation of Pearl's theory of causality having described the theory's key features in section 2.0 of the previous chapter. I set out my case by picking up the thread I began in section 2.1 and develop an interpretive line that remains consistent with the principles of causal modelling outlined in chapter 1. The latter is important since, when exegetical issues arise, it is those principles to which I will turn for guidance. The principal idea of the interpretation I develop throughout this chapter is that Pearl offers a 'two-part' theory of causality. Though I will seek a precise statement of this idea below, the basic claim runs as follows. It is my view that Pearl's theory of causality consists of both an account of causal modelling in science and an account of causal mechanisms as objective constraints on physical processes. Identify the former with what I will call 'Pearl's regimentation of causal concepts for scientists' and identify the latter with what I will call 'Pearl's objective account'. I see the former as a regimentation of a limited set of causal concepts and a set of (largely formal) procedures for their application. These features make up the first part of Pearl's theory. The 'objective account' is a partial analysis of causal mechanisms, which I will later suggest is to be identified with a nomic account of physical processes. The objective account makes up the second part of Pearl's theory. In the final section of this chapter I discuss the question of model justification.

3.1 Pearl's Two-Part Theory of Causation...

Pearl (2000a) claims in his prefatory remarks that the central aim of many studies in the physical, behavioural, social, and biological sciences is the elucidation of cause-effect relationships among variables or events, but that the elucidation undertaken by these sciences has been without the appropriate methodology for extracting causal relationships from data (Pearl 2000a: xiii)[130]. Part of the problem, Pearl claims, has been the lack of a clear semantics for causal talk and the lack of the requisite mathematical machinery suitable for manipulating causal information and identifying and estimating the strength of cause-effect relationships. In other words, Pearl suggests that the existing scattered and disputed methodologies for identifying causes from data need to undergo a process of 'mathematization'[131]. *Causality* is the result of Pearl's sustained effort to

> ...[emphasise] practical methods for elucidating potentially causal relationships from data, deriving causal relationships from combinations of knowledge and data, predicting the effects of actions and policies, evaluating explanations for observed events and scenarios, and–more generally–identifying and explicating the assumptions needed for substantiating causal claims (Pearl 2000a: xiii).

Although Pearl intends his theory to have a wide appeal his primary focus is on those disciplines that attempt to model causal processes such as economics, epidemiology, sociology, and psychology[132]. As I have pointed out above, Pearl considers that causal modelling within these disciplines functions within a specific framework. Pearl's take on this framework involves assuming that nature possesses (often unobservable but) stable causal mechanisms, which it is the task of the causal modeller to discern from

---

[130] These claims are not extended to the world of micro-phenomena.
[131] For discussion see Pearl (2002c, 2003a).
[132] This collection of disciplines has been grouped together and labelled by some as 'engineering disciplines' and thus as applied sciences. Pearl does not extend his theory to the marine sciences or

available observations with or without the aid of some (prior) causal knowledge[133]. This somewhat pragmatic view of causal discovery in science takes care to recognise the situation and requirements of the individual, person, or group of persons who wish to discover or reason about natural causal relationships. I have illustrated many of these commitments in chapter 1.

I suggest that what Pearl (2000a) does when articulating this theory is propound a view of causality that straddles two parts. It is my view that Pearl's theory of causality is the conjunction of an account of causal modelling in the applied sciences and an account of the natural properties assumed to exist by those sciences. The first part of the theory–Pearl's account of causal modelling–resembles a conceptual analysis but isn't one and the second part of the theory–Pearl's objective account–resembles an empirical analysis of the causal relation but isn't complete[134]. In making the suggestion I do not intend to assert this view of causality is in fact the view held by Pearl or the view explicitly put forward by Pearl (2000a). One will not find this view listed in the index or the table of contents. Instead, the point is interpretive. It is what I take Pearl's theory to amount to when examined from a philosophical perspective.

Nevertheless, I claim there is strong support for my interpretation throughout Pearl (2000a) and I detail much of it in section 3.3 below. But first there are several initial points worth making to motivate the interpretation. First, for Pearl (2000a), humans 'possess', 'store', and 'process' causal knowledge and maintain the ability, in certain situations, to observe causal relationships directly and pass on causal information through language (Pearl 2000a: 22, 252-253). To possess such capacities requires that humans are competent users of 'causal talk.' For Pearl (2000a) the competency is increased and advanced, in the context of science, with the aid of (his theory's)

---

related areas in biology and ecology although some have made a partial attempt to do so. See, for instance, Shipley (2000).

[133] That is, the modelling task typically takes place under conditions of either unknown structure and full observability or unknown structure and partial observability.

[134] By 'empirical analysis' I mean following Salmon (1984) an analysis that aims to articulate what causation is as a contingent fact. I discuss the notion of conceptual analysis further below.

formalisation. So, Pearl's theory is a 'mathematization' of human 'causal talk'[135]. But, the fact that Pearl finds it necessary to detail an informal account of the theory's formalism demonstrates the theory amounts to more than just formalism. Put simply, it is my view that the informal account amounts to the explicit specification and regimentation of several key causal concepts beside a set of procedures for their application in non-experimental science[136]. But, since Pearl (2000a) does more than regiment and formalise causal concepts, these are just one component of Pearl's overall theory.

In the attempt to corral causal concepts and provide a formal calculus to aid scientific discovery Pearl is compelled to talk of such things as 'nature's mechanisms,' 'physical laws' and 'objective constraints'. For instance, recall that Pearl believes

> [t]he world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behaviour of a relatively small group of variables (Pearl 2000a: 223).

And further that 'mechanism' is fancy talk for 'law' (Pearl 2000a: 239), and that causal relationships are "...the fundamental building blocks both of physical reality and of human understanding of that reality..." (Pearl 2000a: xiii-xiv), and hence are to be thought of as "ontological, describing objective physical constraints in our world..." (Pearl 2000a: 25)[137]. These remarks are principally ontic in nature and so mark a break from Pearl's formalism and regimentation of causal concepts. There is, therefore, good reason to think that Pearl (2000a) propounds a theory of causality in two parts.

---

[135] For discussion of this issue see Pearl (2000a: 22; 96-97; 135).

[136] See for instance Pearl's claim that "... compliance with human intuition has been the ultimate criterion of adequacy in every philosophical study of causation..." (Pearl 2000a: 26 n12). Compare Glymour's remarks that the causal modelling project, instead of "...'analyzing' the 'concept' of causation, give[s] axiomatic characterisations of the assumptions implicit in large segments of practice" (Glymour 1997: 317).

[137] See also related comments on pages 63, 104, 199, 202, 204, 219, 226, 228, 244, 250, 252 of Pearl (2000a).

It is natural at this point to question how the two parts of the theory are related to each other. As I mentioned above, it is my view that one should see the two parts as two tiers of the one theory. Hence, it is my view that Pearl's theory has a hierarchical structure. The stated aims of Pearl's theory are instructive here. As I mentioned above, Pearl states that his principal aim is to provide formal tools to the applied sciences to aid in the project of causal discovery. Pearl does not attempt to theorise causal discovery for all of science and, as such, does not mean for the theory to have a universal scope[138]. Furthermore, it is my view that it is the first part of Pearl's theory that is designed to aid in the clarification of causal expressions and the drawing of sound causal inferences and that the second part contains statements pertaining to the nature of mechanisms and physical processes. But, be this as it may, I do not see that either part is defined in terms of the other in any straightforward sense. Hence, I do not think that one part is or should be *reducible* to the other. Moreover, I do not consider that the set of equations of a model specify a distinct collection of operations that are to be identified with a set of worldly causal properties. Nor do I think the properties of the processes of the second part (whatever these properties turn out to be) are logically entailed by the conditions on the first part or vice versa, and, even if they were, the relation would not be isomorphic since the first part of Pearl's theory may be consistent with a number of distinctly different second part properties. As such, if models based on the first part fail to refer they are not thereby rendered meaningless. Therefore, it is not part of my interpretation that Pearl's theory is operationalist.

Several of Pearl's comments make it tempting to draw a parallel between Pearl's theory and Hume's approach to the study of causality. Some defend the view that Hume essentially offered a two-part analysis of causality consisting of both an analysis of the causal relation–Hume's so-called 'regularity analysis'–and a distinct analysis of our conception of causation–Hume's conceptual analysis[139]. Hence, the temptation is greater still, given my suggestion that Pearl (2000a) offers a two-part

[138] Pearl draws the line at the border between macro phenomena and micro phenomena.

theory of causality, since one might assimilate the first part of Pearl's theory with Hume's conceptual analysis, and the second part with Hume's regularity analysis[140]. Indeed, several comments throughout Pearl (2000a) are suggestive of such a parallel[141]. But, despite Pearl's numerous comments about Hume's view of causation, the orientation of Pearl's analysis is significantly different from Hume's. In short, Pearl is, on my interpretation, attempting to provide science with better tools for the expression, identification, and estimation of natural causal relationships. Hume, by many accounts, attempted to explain away the very possibility of achieving that goal. Hence, it would be inappropriate to identify the two-parts of Pearl's theory with the (essentially) two-part account of causality given by Hume.

Finally, I mentioned above that the first part of Pearl's theory–Pearl's regimentation–looks like a conceptual analysis but isn't one. This needs some elaboration. There are several different views of conceptual analysis. According to one recent statement conceptual analysis aims to make explicit our folk intuitions about a given topic through the identification and cataloguing of how individuals classify possibilities pertaining to that given topic (Jackson 1998: 31-33). Other notable accounts are due to Strawson (1959), Ryle (1971), Austin (1961), and Grice (1989). At several points throughout Pearl (2000a) Pearl speaks of explicating and protecting human causal intuitions. Moreover, Pearl (2000a) discusses what our concept of causality amounts to as part of a broader project aimed at clarifying communication between groups of scientists. It is tempting to see the sum of these comments and the project of clarification as a form of conceptual analysis of causality.

It is true that Pearl engages in an attempt to make causal intuitions explicit. But, the intuitions Pearl explicates are not really those of the folk (unless one means by folk scientific folk). Pearl's explication occurs within the context of scientific discovery

---

[139] This distinction is drawn in Dowe (2000: 1-2). See also Beauchamp and Rosenberg (1981: 285-290).
[140] Or if one prefers, Hume's counterfactual analysis.
[141] See for example, Pearl (2000: 41, 228, 238, 249, 236, 243).

and so involves several components that stretch folk conceptions of causality. It is also true that Pearl is attentive to the ways in which scientists consider their possibility space when hypothesising. But, again, it would be drawing a long bow to suggest that what a scientist considers possible stems entirely from commonly held intuitions about causality. These considerations count against the view that Pearl (2000a) offers a conceptual analysis of causality. But the main reason against concluding that Pearl is engaged in conceptual analysis is that Pearl does not say that the intuitions he aims to protect and encode are all there is to our concept of causality (Pearl 2000a: 26-27, 220, 257). Hence, it is not appropriate to claim Pearl's project is one of conceptual analysis. Instead, Pearl engages in a practice of identifying, grouping, and constructing procedures around several pre-existing causal concepts as part of his attempt to make reasoning about a limited suit of empirical problems tractable. It is for these reasons I call the first part of Pearl's theory a regimentation of causal concepts for scientists. But, having said that, Pearl does express the view that his account of causal modelling aims at least in part to encode and perhaps extend our understanding of causality, a goal towards which Strawson (1959) aspired also. I conclude by highlighting the fact that the regimentation that takes place on the first part does have the effect of delimiting what may count as a causal relationship on the second part, and that, since the second part places limits upon what counts as a mechanism, it also delimits what form a model may take on the first part[142].

[142] Note however, that each tier stands or falls independently of the other. Since it is not part of Pearl's formal approach to causal modelling that the world actually have any specific causal features, the first tier would not fall on arguments to the effect that the world actually possesses or fail to possess some specific causal features. Instead these issues might weigh against the formalism's measures of success. Likewise, that the world actually possesses some specific causal features does not logically require that a formal approach to its modelling have or lack the features of Pearl's account.

3.2 ...and its Two Riders

Many details need to be teased out and clarified from the picture just presented, especially since the content and purpose of the two parts may seem to many to be in tension. The forgoing picture is only cursory. I offer a more detailed examination directly. But before turning to that task, a further issue begs attention. Pearl apparently places two riders on the claims I have attributed to him regarding the nature of his theory of causal modelling and the world in which that modelling takes place. The first rider states that causality exists only relative to human agency (Pearl 2000a: 349-350). The second rider states that causal models must be consistent with the symmetry of physical laws (Pearl 2000a: 349, 223-228).

The first rider applies for the following reason. Pearl thinks causality is agent relative in as much as the causal conceptions, explanations, and causal talk of humans respects both a temporal direction and causal asymmetry the basis for which is not a feature of physical reality. For instance, Pearl asserts:

> [...] certain patterns of dependency, which are totally devoid of temporal information, are conceptually characteristic of certain causal directionalities and not others. Reichenbach (1956) suggested that this directionality is a characteristic of Nature, reflective of the temporal asymmetries associated with the second law of thermodynamics. [...] We offer a more subjective explanation, attributing the directionality to choice of language and to certain assumptions [such as stability] prevalent in scientific induction (Pearl 2000a: 43).
>
> [Furthermore,] [...] scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece *in*–namely the focus of investigation. [...] The choice of *ins* and *outs* creates asymmetry in the way we look at things,

> and it is this asymmetry that permits us to talk about 'outside intervention' and hence about causality and cause-effect directionality (Pearl 2000a: 350).

Pearl offers the following line of reasoning for this position. First, what Pearl means by 'temporal asymmetry' is captured by the requirement that, in all cases, causes precede their effects in time[143]. The asymmetry is salient for Pearl since he claims temporal information is one of the best guides to causal structure but that, even without temporal information, it is possible to identify causes from data (Pearl 2000a: 42). As such, Pearl believes we are owed a two-fold explanation. First, how is it that statistical information happens, as it turns out, to respect a temporal asymmetry, and, second, why do our causal judgements coincide with such a temporal asymmetry. Pearl thinks the appropriate explanation of these facts must tell us how causal directionality can be discerned from bare statistical information, and why the causes thus discerned coincide with the (apparent) direction of time.

Pearl responds by first providing a formal definition of statistical asymmetry via the notion of 'statistical time':

> Statistical Time
> *Given an empirical distribution P, a statistical time for P is any ordering of the variables that agrees with at least one minimal causal structure consistent with P*[144].

Pearl then formalises an expression of the notion of 'temporal bias':

> Temporal Bias
> *In most natural phenomena, the physical time coincides with at least one statistical time*[145].

---

[143] The notion of temporal asymmetry is of course far broader than this. It is so broad that I will not attempt a summary. See Price (1996) and Savitt (1995) for a variety of contemporary and historical views of temporal asymmetry. I remind the reader that Pearl does not suppose his theory to be universal in scope.

Recall that a causal structure over a set of variables V is a DAG such that each node corresponds to a distinct element in V and where each link between two nodes represents direct functional relationships among the corresponding variables (Pearl 2000a: 44). Further recall that, granted the disturbance terms associated with a causal model (Causal structure plus parameters) are independent, the model satisfies the Markov assumption. Pearl considers the Markov assumption to be underpinned by Reichenbach's Principle of Common Cause (PCC). In a generic form the PCC states that, for any non-accidental correlation between two distinct events, which are not themselves causally related to each other, there exists a common cause that screens off their correlation and thus renders them probabilistically independent (Reichenbach 1956: 157-160). But, according to Pearl, a causal structure may admit various statistical times (Pearl 2000a: 58-59). Hence, it is possible to show that any statistical time can run contrary to physical time just by choosing a different linguistic representation of causal structures of observed distributions (Pearl 2000a: 59).

Pearl thinks that solving the problem requires one to recognise the fact that "[...] the consistent agreement between physical and statistical times is a byproduct of the human choice of linguistic primitives and not a feature of physical reality" (Pearl 2000a: 59). All that remains to be explained is the particular choice of linguistic primitives. Pearl speculates that the choice may have something to do with selection pressures to facilitate predictions for future rather than past events (Pearl 2000a: 59)[146].

Having dealt with temporal direction, causal asymmetry remains. The task here, according to Pearl, is to explain our choice of causes as causes instead of effects. And, again, the issue is salient for Pearl since Pearl holds that causal models admit

---

[144] This definition appears on p 58 of Pearl (2000a).
[145] The definition of this conjecture appears on p 59 of Pearl (2000a).
[146] It would follow from this that Pearl takes the PCC to be a regulative epistemic principle rather than an empirical or metaphysical one. I disagree that reducing the problem at hand to choice of linguistic primitives under selection pressure is adequate. In fact, I believe such a reduction merely shifts the problem to another level. It is curious that Pearl cites Price (1996) but does not mention his solution.

relations of asymmetric dependency between variables that are causes and variables that are effects, and that in some instances this causal direction may be discerned from the topological features of a distribution that is itself devoid of a temporal ordering or indexation[147].

Pearl (2000a) asserts that causal ordering (of a model) ensues from one of two assumptions. First, the choice made by the investigator to partition the variables (events) into background and endogenous sets, and second, the overall configuration of mechanisms in the model (Pearl 2000a: 226-227)[148]. In essence, the identification of causes requires locating those variables within the distribution that remain invariant to local actions or surgeries. This is so since manipulations performed on the endogenous variables of a model are (in some cases) able to reveal the dependence relations present within the distribution. However, Pearl does not think that the asymmetry defined for a model is a constituent of nature, and instead he chooses to explain the asymmetry of cause to effect as an artefact of the modeller's interests. Pearl discusses the issue via a response to Russell's assertion that causation is inconsistent with the teachings of modern physics[149]. Pearl's response to Russell admits that:

> [...] the equations of physics are indeed symmetrical, but when we compare the phrases 'A causes B' versus 'B causes A,' we are not talking about a single set of equations. Rather, we are comparing two world models, represented by two different sets of equations: one in which the equation for A is surgically removed; the other where the equation for B is removed. Russell would probably stop us at this point and ask: "How can you talk about two world models when in fact there is only one world model, given by all the equations of physics put together?" The answer is: yes. If you wish to include the entire universe in the model, causality disappears because interventions

[147] See Pearl (2000a: 226-228) for detailed discussion.
[148] It follows that particular variables or events may be causes or effects depending on the model they are part of. The two assumptions Pearl speaks of originate with the work of Simon (1953).

> disappear - the manipulator and the manipulated loose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece *in* - namely, the focus of investigation. The rest of the universe is then considered *out* or background and is summarised by what we call boundary conditions. This choice of *ins* and *outs* creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about 'outside intervention' and hence about causality and cause-effect directionality (Pearl 2000a: 349-350).[150]

Hence, the variables that are causes rather than effects result from the investigator's choice to partition the variable space into exogenous and endogenous sets. In the present context the variable space coincides with the investigator's observations and the partition amounts to an interest in ascertaining what happens to one set when the other is altered in some controlled fashion. However, the key point is that the asymmetry that results from the partitioning is for Pearl not a discovery but an imposition. The asymmetry does not exist beyond our conceptual apparatus.

The second rider also relates to the direction of causality. Recall that Pearl places a second rider on his theory to the effect that the asymmetry of equations in a causal model is consistent with the symmetry of the equations common to modern physics. Pearl is concerned that the asymmetry of causal models is inconsistent with the apparent lack of asymmetry in physical equations. Against that case Pearl asserts:

---

[149] It is arguable that Pearl has misunderstood Russell's assertion that causation is inconsistent with modern physics. This will be of some minor importance below.

[150] Given this statement, some may wonder why I bother to lumber Pearl with empirical analysis, since Pearl apparently thinks that causation does not exist beyond human conceptions. However, the point ignores the fact that Pearl's theory is supposed to aid scientists in investigating the world, not human conceptual spaces. Pearl is asserting that scientists are forced to stop talking cause and effect just in case we are disallowed boundaries. In this case I read Pearl to be claiming that causality ceases to exist, not because there are no objective physical processes but instead because we cannot discern directionality in cases where we are banned from making effective manipulations. In effect, Pearl is pointing out that modelling (and presumably most science as it is presently practiced) would be pointless in such instances.

> The asymmetry that characterises causal relationships in no way conflicts with the symmetry of physical equations. By saying that '*X* causes *Y* and *Y* does not cause *X*,' we mean to say that changing a mechanism in which *X* is normally the dependent variable has a different effect on the world than changing a mechanism in which *Y* is normally the dependent variable. Because two separate mechanisms are involved, the statement stands in perfect harmony with the symmetry we find in the equations of physics (Pearl 2000a: 228).

Exactly what Pearl has in mind here is difficult to ascertain. There is a sense in which the symmetry of equations commonly utilised in physics (such as differential equations) is a matter of mathematical form. For instance, Newton's equations are symmetric in this sense. There is a distinct sense in which one means by symmetry in physics that processes permissible by an equation in one temporal direction are equally permissible in the opposite temporal direction (Price 1996: 116). Indeed, Pearl's talk of the symmetry of equations does seem to imply something more fundamental than mere symmetry of representation. To the extent that one might correctly read Pearl (2000a) to mean either, Pearl's assertions regarding symmetry in physics are ambiguous[151]. The following comment by Pearl (2000a) adds some clarity. When Pearl (2000a) asserts that the relationship between dependent and independent variables in a causal model involves two distinct mechanisms he means to say the relationship is not governed by a single set of equations:

> Rather, we are comparing two world models, represented by two different sets of equations: one in which the equation for A is surgically removed; the other where the equation for B is removed (Pearl 2000a: 349).

So, any given causal model is merely half of what one may think of as the 'complete' model for a given set of relationships. The 'complete' model is complete in the sense

[151] I note that Pearl does not attempt to spell out which symmetry of several available symmetries he has in mind.

that it coincides with the (appropriate set of) physical equations describing said relationships. According to Pearl (2000a) one may, therefore, break symmetry into a pair of asymmetries or convert a pair of asymmetries into symmetry. Hence, Pearl thinks there is no conflict between the asymmetry of causal models and the symmetry of physical equations because he has a procedure for breaking symmetries into asymmetries and vice versa. Pearl (2000a) says nothing more decisive on the matter and declines to make use of the definitions of asymmetry provided by Price (1996) and so stands the second rider.

But to my mind the ambiguity remains. It is unclear whether Pearl (2000a) merely aims to maintain consistency with the form of representation chosen by modern physics or aims to demonstrate that the asymmetry of causal models does not conflict with the symmetry of physical laws. My reading leans towards the former. Note, however, that if the latter is Pearl's intention, making the case is not a straightforward matter for several reasons. For instance, a causal mechanism is, for Pearl, to be identified with objective physical constraints. Presumably, if physical constraints delimit objective physical processes, then if there are two such mechanisms for each symmetric constraint there ought to be two (asymmetric) physical processes to match. Or, put another way, the existence of pairs of physical processes (where one affects past events only and the other affects future events only) entails there are matching pairs of physical constraints. This result–being prone to problems of overdetermination–would then place considerable weight on Pearl's explanation of temporal direction[152].

[152] I presume the assumption is also in conflict with modern physics.

3.3 Modelling Causal Processes: A closer look at the 'two-part' interpretation

The forgoing sections set out a framework within which to understand the nature of Pearl's theory of causality, including the theory's two riders. I can now proceed to consider what the theory says about causality in greater detail. How complete these considerations can be is, of course, limited by the level of detail present in Pearl (2000a) and related sources.

Not surprisingly, as there are two parts to the theory each part makes its own distinct contribution to the theory. The first part principally offers an account of causal modelling, and so, what causality is on this part directly concerns the nature of causal inference. From above, I found the account of causal modelling to consist of a regimentation and a formalisation of causality. The salient question to pose on this tier then should be more specific than simply 'What is causality?' Instead, it is more appropriate to ask 'What is a *causal* inference?' The aim of posing the question is to elicit an answer that lets one know what makes a given inference *causal* rather than (say) statistical.

Unfortunately, Pearl's answer to this rephrased question is akin to the answer one might expect to receive after asking an appropriately qualified person 'what is a good move in a game of Chess?,' namely, multi-faceted and rather longwinded. Hitchcock offers the following summary:

> Most philosophers talk as though there is one specific relation–causation–that is the target of philosophical inquiry. In [Pearl (2000a)], one finds definitions of causal effect, causal relevance, total effect, direct effect, actual cause, contributing cause, and so on; the 'Causality' of Pearl's title does not name some specific relation, but rather an entire subject matter. Pearl's concept of actual causation comes closest to the notion of 'token causation' that most philosophers take to be central. It is telling that this concept appears only in

> Pearl's final chapter; the concept is not needed for Pearl's treatment of rational deliberation, counterfactuals, experimental methodology, and so on (Hitchcock 2001: 640).

One lesson to take away from Hitchcock's summary is that the subject matter addressed by Pearl's theory is rather diverse, and so, presenting an exhaustive answer to the question at hand would necessitate covering considerable ground and involve unnecessary repetition of detail already presented in chapters 1 and 2[153]. Nonetheless, some headway can be made. One claim I asserted earlier was to the effect that Pearl's account (of causal modelling) really begins with the causal Bayesian Network and its associated interpretation[154]. As such, and so as to avoid repetition, I will use that object as the foil for a less evasive answer. The consequence of focussing on causal Bayesian Networks in this way is that the question must be rephrased to suit. The appropriate formulation is 'What (on the first tier of Pearl's theory) does the term 'causal' mean?' or, more succinctly; 'What does Pearl mean by the term 'causal' in 'causal Bayesian network' and 'causal model'?' I do not believe the answer to this question is obvious or trivial. I reject out of hand answers taking the line that what the term 'causal' means here amounts to nothing more than a label placed on the output of iterations of Pearl's formalism. It is evident that Pearl rejects this line also.

Pearl thinks that a Bayesian network becomes a 'causal' Bayesian network when it is furnished with a causal interpretation. That is, Bayesian networks are causal networks when interpreted

> ...as a system of processes, one per family, that could account for the generation of the observed data. [...] [Such that] each parent-child

---

[153] I again draw the reader's attention to the disparity, introduced in chapter 1, that exists between scientific orientations towards issues of causality and the orientations of the philosophy of science on this matter. Pearl's answer might be seen to be longwinded because he appears to be answering a 'how' question (i.e. 'How does one draw valid causal inferences?') rather than attempting to answer a 'what' or 'why' question (i.e. 'Why is this inference causal?' or even 'Why is this causal inference valid?'

[154] The reader should include the updated form of the causal Bayesian Net (Causal Models) as the referent in the following discussion.

> relationship in the network represents a stable and autonomous physical mechanism [...] [and where] it is conceivable to change one such relationship *without* changing the others (Pearl 2000a: 21, 22 original emphasis).

When a Bayesian Network is given a causal rather than an associational interpretation the missing links of its graph represent the absence of causal connections rather than statistical independence (Pearl 2000a: 141). In other words, as I discussed in section 2.0.3, Pearl holds that the meaning of the term 'causal', where it appears in the elaboration of his account of causal modelling, is read into the account's formal features. Were one to define the term 'cause', based solely on the formal features of the account (i.e. sections 2.0.1 and 2.0.2 but minus section 2.0.3 above) the term and its cognates become primitives. The comments are further suggestive of the fact that Pearl intends the formal component to be consistent with his view of physical constraints, which are yet to be dealt with[155]. In any case, these and other comments see to it that for Pearl the term 'cause' is defined as the conjunction of the following four conditions:

> 1. Determinism: the relation between causes and effects is mediated by an asymmetric deterministic structure.
> 2. The Causal Markov Condition: a deterministic structure is Markovian in the sense that each variable X that is part of such a structure is independent of all its non-descendants, given its parents[156].
> 3. Modularity: each term that forms part of a causal structure exhibits the possibility of alteration (via interventions) without that alteration affecting any other terms in the structure.
> 4. Minimality: the simplest causal structure that fits the data is the preferred structure.

---

[155] Such commitments would surface in answer to the question 'Why have you constructed the system like that $\nearrow$?' Recall Pearl's comments in the postscript to chapter 3 Pearl (2000a: 104-105).
[156] The CMC is often accompanied by a sufficiency condition. Broadly speaking, a sufficiency condition states that a set of variables V is causally sufficient for a population just in case in the population any common cause of two or more variables in V is also in V or has the same value for all the units in the population (Berkovitz 2002: 259-260). See also Scheines (1997).

5. Stability: a distribution is stable when the independencies that occur in it are the result of structure rather than chance.

The conjunction of these conditions may also be thought of as the licence for *causal* inference, where, as one and another conjunct is removed from the definition, so it becomes more difficult to speak meaningfully of inferences as *causal*. More importantly, Pearl admits that the conjunction of these conditions presupposes either one of two conceptual categories. The two categories are *functional relationships* represented by structural equations or *manipulations* that substitute one equation for another under the guidance of the calculus of interventions (Pearl 2000a: 30 n17)[157]. For Pearl these two categories are 'models for causation' in the sense that they offer a context within which to understand each condition. In effect, the conceptualisation of causality as either functional mechanisms or manipulations coincides with the purpose of the informal semantics of section 2.0.3, adding to the primitive notions of the formal components a causal meaning. These notions plus the conjunction of the five conditions stated above and Pearl's formal machinery are what I call Pearl's regimentation.

It is worth taking a brief look at each condition in turn before moving on to discuss the second tier. Condition 1 appears to be claiming nothing stronger than the output or dependent variables of structural equations are fully specified as a function of the variables on the l.h.s. of such equations[158]. The relevant meaning of 'determinism' is provided by the functions of causal models and not the physical constraints or processes such functions may serve to represent. Condition 2 follows, according to

[157] See also discussion on the role of mechanisms in defining causality in Pearl (2000a: 25).

[158] I mean this in the sense that, were the values of all variables and parameters that figure in the r.h.s. of the structural equation known, then so too would the value of the variable on the l.h.s. Recall, and further that the equality sign of structural equations does not function as does the algebraic equality sign. The former is closer to dependence than equality. Pearl makes explicit reference to 'deterministic functions' in his preface. But there the statement would appear to suggest reference to such functions merely as representations of actual deterministic physical processes. As such I take those claims to be part of the second tier to Pearl's account. The deterministic nature of functions is in contrast taken to refer to properties of causal models and as such belongs on the first tier of the account.

Pearl, from two causal assumptions; minimality (condition 4 above) and Reichenbach's (1956) common cause assumption:

> if any two variables are dependent, then one is the cause of the other or there is a third variable causing both (Pearl 2000a: 30).

I discuss both assumptions in greater detail below. The condition of modularity pertains to the possibility of altering a mechanism featured in a model without thereby altering any other distinct mechanisms in the model. Minimality is a property of a causal structure just when there exists no distinct alternative structure with less observed parameters capable of reproducing equivalent independencies. Stability requires that the configuration of mechanisms in a model produce all and only those independencies present in the distribution associated with the model. Pearl's concepts of mechanism and manipulation are important here. The modularity condition is the modelling analogue of the concept of the *autonomy* of physical mechanisms[159]. According to the concept of autonomy, bona fide physical mechanisms remain invariant to certain forms of intervention. Pearl provides the following summary:

> A causal model is not just another scheme of encoding probability distribution (sic) through a set of parameters. When we come to define mathematical objects such as causal models, we must ensure that the definition captures the distinct ways in which these objects are being used and conceptualised. The distinctive feature of causal models is that each variable is determined by a set of other variables through a relationship (called 'mechanism') that remains *invariant* when those other variables are subjected to external influences. Only by virtue of this invariance do causal models allows us to predict the effects of changes and interventions, capitalising on the locality of such changes (Pearl 2000a: 63 original emphasis).

[159] It is with the notion of autonomy that counterfactuals are introduced.

A simple way to think of a causal model then is as a formal representation of a system of autonomous mechanisms explicitly constructed to meet the five conditions set out above so as to give the model every opportunity to licence inferences that coincide with the scientist's concept of causal relationships (and causal knowledge or evidence). Hence, what the term 'cause' means, is expressed through the five conditions cited above in the context of a functional mechanism that admits alteration. In short, a model that embodies all five conditions can be seen to express a set of causal relationships[160]. The idea is that such models may be used to guide inferences, which track pertinent physical relationships in the part of the world the scientist has under investigation. The latter notion moves the discussion onto the second part of Pearl's theory.

The question of what causality amounts to on the second tier is far from obvious. There are at least two reasons for this. First, Pearl's (2000a) comments on the matter are sparse and invariably lack the depth necessary to assemble a definition. Second, it is common that one happens across comments that at first appear to refer directly to the nature of the relevant physical systems or processes but, on closer examination turn out to be statements made either; (i) in the context of the formal component of Pearl's account or; (ii) to features of human psychology thought (by Pearl) to be relevant to learning and storing causal information. In such instances one is struck by the fact that–contrary to first appearances–neither of these sorts of statements shed much light on the subject. The result is that Pearl's second tier commitments cannot be straightforwardly and un-problematically read off his interpretation of a formal system–let alone his views of human psychology. Indeed, despite the existence of numerous claims made throughout Pearl (2000a), which I claim commit Pearl to a second tier, none prove to be particularly suggestive of specific details. Even so, some of these comments do seem to travel some small way towards revealing Pearl's

[160] I note in passing that missing from Pearl's first tier account are conditions relating to spatial and temporal contiguity. These conditions are assumed under the CMC and sufficiency condition. See below for discussion. See also Pearl (2000a: 63-64). I reiterate the point that some use Markovian models without attaching any causal interpretation to them whatsoever.

second tier commitments and so deserve closer examination. For instance, the following comment appears in a short discussion of mechanisms:

> The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behaviour of a relatively small group of variables (Pearl 2000a: 223).

On my view, the first part of this comment–that pertaining to the world–is to be identified with the first tier of Pearl's account rather than the second. This is so principally because the properties of autonomy and invariance are defined in terms of–and hence are relativised to–causal models and not to actual causal processes. I take the reference to the 'world' in this passage to denote the modeller's world[161]. However, the second part of the comment claims that each of these properties *corresponds* to a physical process[162]. What does Pearl mean by this?

In a remark about the difference between probabilistic and causal relationships Pearl asserts that

> [...] causal relationships are *ontological*, describing objective physical constraints in our world [...] (Pearl 2000a: 25 original emphasis).

And further that structural assumptions (such as the CMC) are

> [...] adopted as a useful abstraction of the underlying physical processes because such processes are too detailed to be of practical use (Pearl 2000a: 61).

---

[161] See above for a definition and discussion of the causal modeller's world; especially section 1.2. For further comments supporting this interpretation see Pearl (2000a: 68, 219). But note that comments made later on p 250 count against this interpretation. On closer examination I suggest a slide has occurred from defining invariance and autonomy in Section 2.9.1 explicitly as properties of models to defining these properties in terms of physical processes. I prefer the earlier treatment.

[162] Physical processes are not to be confused or identified with active causal processes in Pearl's account. The latter are, in effect, a type of counterfactual dependency between variables in a causal model within a specified context. See Pearl (2000a: 318) and section 3 of Halpern and Pearl (2001a).

These latter comments add something to Pearl's notion of physical processes. Together the comments suggest that Pearl means to tie the notion of law-like mechanisms or 'constraints' together with physical processes. The mechanisms are 'law-like' in the sense that the relationship that obtains between the variables or events of a physical process bound by a mechanism is invariant to some specific class of perturbations. Pearl's identification of causal relationships with law-governed physical processes is suggestive of a 'nomic' account of causation. The general features of nomic accounts of causation are, therefore, worthy of some discussion as they may shed some light on Pearl's second tier commitments[163].

To adopt a nomic account of causality would be (minimally) to take the view that when two quantities or events are truly *causally* related then one quantity or event *necessitates* the other in a 'law-like' fashion (i.e. almost without exception) (Psillos 2002: 169). One example of a nomic account of causation is that offered by Armstrong (1997). Put simply, Armstrong is committed to the view that two event (types) are related by law iff there is a relation of nomic necessitation–written N(F,G)–between the properties (which are universals for Armstrong) F-ness and G-ness where all F's are G's (Psillos 2002: 163), and where, for Armstrong (1997), the relation of Necessitation that exists between the properties (universals) F and G is the causal relation. That is, the putative causal relation that exists between two singular events is in fact a causal relation just in case it is the relation that exists between two universals of which the singular events are tokens[164].

Adding the sort of nomic account of causation that Armstrong elaborates to Pearl's theory would have the following result. Pearl's regimentation remains a formal

---

[163] I note in passing that Pearl does not address issues concerning the reduction of causal or law-like relationships in the special sciences to the laws of physics raised in Putnam (1973) Fodor (1974).
[164] On Armstrong's view then causation requires the instantiation of laws. This is not, of course Pearl's view, but the question naturally arises as to whether Pearl's reference to laws is Humean or non-Humean in spirit. In lieu of further investigation I conjecture that Pearl would not be forced to commit either way. This is another issue where the 'how?' of causal discovery in science displaces (the importance of) the 'why?' Pearl is not (obviously) doing metaphysics.

approach to causal discovery but where 'causal relationships' are those of the nomic variety. The role of Pearl's objective account is then to offer a general set of conditions defining such nomic relationships. This would be an attractive interpretation of Pearl's account but for the following difficulties. First, Pearl allows the granularity of causal relationships to be decided by the investigator on a case-by-case basis. This is what Pearl means when he speaks of abstracting away from actual physical processes. As such it could not be the case that the events or quantities of Pearl's account coincide with the particular matters of fact or states of affairs required by a nomic account such as Armstrong's. Second, on Pearl's account truth is relativised to a model rather than to eternal statements or that-clauses and so on. Third, Pearl does not appear to expound or endorse (or require) an account of natural laws of the sort required by a nomic account of causality such as Armstrong's[165]. But, even so, I think the nomic account is the natural (philosophical) interpretation of Pearl's comments on objective constraints. Perhaps it is possible to alter the objectionable features of the nomic account so that it may be allied to the second part of Pearl's theory. I shall consider this line of thought briefly.

Altering the nomic account will involve revisiting Glennan's (2000a, 2002) account of mechanical models from section 1.1. Recall that I described there how Pearl's theory of causality could be given a suitably general philosophical context through a connection with Glennan's account of mechanism. The salient points of Glennan's account are as follows. Glennan (2000a, 2000b, 2002) defines a mechanism underlying a behaviour as a complex system that produces the behaviour through the interaction of a number of parts such that the interactions between the parts of the mechanism can be characterised by direct, invariant, change-relating laws (Glennan 2002: S344). The notion of a mechanism then serves as a basis for Glennan's account of a mechanical model. Importantly, I noted in section 1.1 that Glennan's characterisation of mechanical models has two parts. One part is concerned with the

[165] A variation on the theme of taking the first tier to offer a formalism for causal discovery and the second tier to define the causal relation would be to assert that the second tier defines causality in terms of the transference of conserved quantities along the lines of Dowe (2000). However, there is little or no support for this interpretation either.

description of the *behaviour* of the mechanism, the other part with model's mechanical description. Glennan's point is that the mechanical description describes the internal structure of the mechanism whilst the description of the mechanism's behaviour is an external description. If one considers the behaviour of a mechanism to involve a set of observations, then one may identify external descriptions with causal models and internal descriptions with objective constraints. These two parts of Glennan's account then coincide with the first and second parts of Pearl's theory respectively. Fortifying Pearl's theory with Glennan's account of models endows it with the following features.

The *behaviour* of a system of physical processes is described by the values of the variables of a causal model and the *mechanism* is described by the causal model's structure. Recall that for Pearl, mechanisms constrain processes in the sense that the variables related via a mechanism are invariant to some class of interventions. A causal model then encodes a class of invariant relationships[166]. Drawing the connection between the two parts of Pearl's theory and Glennan's two-part account of mechanism in the way I suggest adds to the former by illustrating the role the second part plays in Pearl's theory. Note, however, that on Glennan's account the relation that obtains between a model and a mechanism is one of 'approximate similarity' in the sense that the description of the mechanism offers only a degree of approximation of reality. That is, on Glennan's account, to make claims about the nature of a mechanism, one constructs a model and asserts that it is similar to a system in nature (Glennan 2000a p 12)[167]. But, what exactly is a system in nature? If one grants the connection between Pearl's theory and Glennan's account of mechanism, then the advance toward Pearl's second part answer depends upon an account of approximate similarity[168].

---

[166] Which Pearl describes with counterfactual sentences. Woodward (2000, 2001, 2002a) explore the notion of constraints as domains of invariance and details a concomitant account of explanation.
[167] I note in passing that there are several other accounts of mechanism besides Glennan's available in the literature. See for instance, Machamer et al. (2000) and Tabery (2004). I prefer Glennan's account since it appears the most consistent with Pearl's account.
[168] I return to the issue of approximate similarity and the relationship between Pearl's and Glennan's view of mechanism when I discuss identification and model justification in section 3.4.

Hence, a frank assessment of the prospects pursued thus far must conclude that one is not carried a great deal closer to understanding what exactly Pearl thinks a physical process is. I would like to suggest that Pearl is committed to the view that causality is in some sense the instantiation of contingent law-like connections between events but the level of detail available does not warrant it[169]. So far then the nature of a 'physical process' remains almost entirely obscure in Pearl's account. However, the significance of Pearl's comments regarding objective constraints dictates that I consider any available avenue around the impasse.

One might think it possible and worthwhile to attempt to draw inferences from the content of the regimentation or from Pearl's views of human psychology to arrive at the commitments Pearl does or ought to have regarding the nature of causal processes. But, since I have claimed above that each part of Pearl's account is independent of the other, this strategy, even if it were informative, is not likely to have any prescriptive weight. Several of Pearl's statements concerning human psychology and causal assumptions are suggestive though, and so I will follow up this line of thought despite reservations.

In regard to the psychology of causal learning Pearl claims that:

> Structural equations and their associated graphs are particularly useful as a means of expressing assumptions about cause-effect relationships. Such assumptions rest on prior experiential knowledge, which–as suggested by ample evidence–is encoded in the human mind in terms of interconnected assemblies of autonomous mechanisms. These mechanisms are thus the building blocks from which judgements about counterfactuals are derived. Structural equations $\{f\}$ and their graphical abstraction $G(M)$ provide direct mappings for these mechanisms and therefore constitute a natural language

[169] This phraseology is from Dowe (2000: 169).

for articulating or verifying causal knowledge or assumptions (Pearl 2000a: 244).

The basic picture Pearl appears to articulate here and elsewhere involves a relation between four things:

1. Human psychological properties.
2. Human knowledge
3. Human causal assumptions.
4. Pearl's formal language.

As far as I can see the relation between the four is as follows. Humans possess a certain psychological capacity or mental structure such that causal information[170], especially counterfactual information, is gathered and stored in the form of autonomous mechanisms. Human *a posteriori* knowledge, especially knowledge of causal relationships, reflects the structure of our psychology. In turn, the structure of our causal knowledge and assumptions is accessible via (i.e. reflected in) and expressed through our causal judgements as these judgements appear in our natural languages. Last, Pearl's account borrows from our psychology its assemblies of autonomous mechanisms and utilises them as primitives in a formal language which, of course, Pearl intends to be a tool for organising and guiding our causal judgements and building knowledge of our environment[171].

Do these relations suggest or rule out any specific features of nature and in so doing move us closer to Pearl's views on physical processes? I cannot say that they do. Even if it were true that Pearl's formalism offered a 'direct mapping' of human mental structures the nature of reality need not thereby fit the specific mould cast for it[172]. In

[170] Gained via observing and interacting with the world.
[171] Where counterfactuals are those endorsed in Pearl's account and causal information is to be contrasted with statistical information.
[172] Although, I am aware it may be the case that Pearl intends to elaborate on a theme familiar from the evolutionary epistemology project. If it were the case that our causal reasoning 'apparatus' was in some coherent sense the result of selection pressures then Pearl might advocate a naturalistic line of

any case, Pearl (2000a) offers no argument for the latter position and the former is controversial. For instance, Pearl's view here is to the effect that human mental structures are governed by or instantiate causal primitives and not (probabilistic) parametric primitives. Such a position takes an inversion of Fisherian epistemology, and utilises it as a psychological foundation. Pearl cites Tversky and Kahneman (1980) as the source of the empirical support for this view. The standout passage that demonstrates Pearl's commitment to such a view of human psychological structure reads as follows:

> ...humans are generally oblivious to rates and proportions (which are transitory) and [instead] they constantly search for causal relations (which are invariant). Once people interpret proportions as causal relations, they continue to process those relations by causal calculus and not by the calculus of proportions. Were our minds governed by the calculus of proportions [...] Simpson's paradox would never have generated the attention that it did (Pearl 2000a: 182)[173].

Although I confess some sympathy for the epistemic equivalent of this view I do not follow Pearl's journey into psychology. Nor do Cosmides and Tooby (1996) who argue that (amongst other things) uncovering evidence (allegedly) supporting the view that humans are poor statistical reasoners, and thus that human psychology is not naturally inclined to probabilistic or proportional reasoning, depends crucially upon the experimental design from which such evidence is thought to arise. They claim to have shown through a series of experiments that people's performance in reasoning with proportions gets better due purely to the way in which the problems are presented. From this fact Cosmides and Tooby (1996) go on to argue the opposite view to that reached by Pearl, namely that people *are* proportional reasoners by

---

reasoning to the effect that the success of our reasoning and the power of our causal explanations commits us to some form of direct realism regarding (some) causal relationships. Since Pearl (2000a) does not explicitly elaborate this line I mention it only to set it aside. But see Pearl (2000a: 59, 253) for supportive comments.

[173] Pearl also cites the work of Waldmann *et al.* (1995) in support of this position. See Pearl (2000a: 60).

nature. What is clear is that the work of Tversky and Kahneman is not the last word on the matter and so Pearl's claims concerning human psychology cannot be decisive[174].

However debates concerning mental content turn out the upshot as I see it for present purposes is as follows. Even if inferences from these components of Pearl's account to the relevant nature of physical processes were coherent and sanctioned (and I find no clear grounds in the literature that establish the validity of such inferences) the basis from which such inferences spring is questionable. Now that I have exhausted the promising avenues I move on to evaluate the regimentation's answer.

Is Pearl's regimentation cogent? Many will not think so. It seems there are ample grounds to challenge each and every conjunct that forms part of Pearl's regimentation. Many problems have been cited across several disciplines, which pertain to either the nature of Pearl's project or specific portions of its formal machinery and underlying assumptions. For example, Heckman (2001) offers a review of modelling and related issues from an econometric standpoint, Dawid (2000), Pratt and Schlaifer (1984), Singpurwalla (2002) and Lemmer (1993) offer a critical statistical perspective, Hedstrom and Swedberg (1998) edit a collection of essays written by sociologists, and Freedman (1997, 2002, 2003), Humphreys (1997), Humphreys and Freedman (1996, 1999) offer critical reviews that focus on graphical models. More recently Cartwright (1995, 2000), Woodward (2001, 2002b), Hausman and Woodward (1999), and Hopkins and Pearl (2003) offer critical review of several formal components specific to Pearl's project as do Moldonado and Greenland (2002) and Cole and Hernan (2002).

Before looking into the detail of the issues raised in the literature it will be useful to introduce a partition between two forms of criticism Pearl's regimentation faces[175].

---

[174] On the broader issue of whether Pearl is referring to cognitive algorithms or culturally conditioned behaviours I suggest Pearl means to refer to the former. For critical discussion of the distinction see Hankinson Nelson (2003: 279).

On the one side of the partition are those criticisms that originate from statisticians, whether applied or otherwise, and on the other side are more thoroughly philosophical criticisms. The two sides of the partition are not mutually exclusive. Some of the criticism offered by statisticians is at least in part philosophical in nature and likewise philosophical criticisms sometimes pertain to statistical technicalities. Even so, I make the partition for three reasons. First, Pearl appears to abide by a similar partition when replying to criticism. Second, identifying the origin of specific criticism allows one the opportunity to respect the intentions of those who offer it, intentions that often reflect specific disciplinary concerns. Third, I wish to focus on criticism offered by philosophers without neglecting altogether criticism offered by statisticians and others.

The criticisms offered by statisticians, or, at least, the criticisms offered on statistical grounds, tend to revolve around one specific point. It is that causality may be adequately dealt with within the boundaries of probability theory; no extra mathematical machinery and no new language is required to deal with causality. Hence, so the claim goes, Pearl's account of what causality *is* is misguided and un-necessary. Less sophisticated versions of this criticism elaborate on the familiar theme that correlation does not equal causation, and that Pearl misunderstands the limited place of causal inference in statistics. Criticism taking this form stems from the worldview according to which causation is a form of correlation between stochastic variables or events. More sophisticated criticisms of Pearl's answer take the view that statistical tools are already available to adequately cope with causal inferences and the identification of causal effects. For instance, Nozer Singpurwalla asserts that 'the calculus of probabilities, endowed with a time dynamic, is indeed the calculus of causality' and seeks to demonstrate this by examples (Pearl 2002b: 210). Pearl believes two points underpin most criticisms of his theory by statisticians. The first is that causal analysis starts with theoretically or judgementally based assumptions for which there exists no support from (frequency based) data. The

[175] I explicitly exclude criticism of the formal component of Pearl's account from the partition. The key characteristic of such criticism is typically the focus on the richness of the semantics. Some claim that

second is that causal analysis, at least as Pearl conceives it, requires an extension to the syntax of the probability calculus (principally with the '$do(x)$' operator). Apparently statisticians find this requirement unacceptable[176]. Other criticisms find that complex systems such as economies can be modelled adequately without recourse to causal models (LeRoy 2002), or that the key components of Pearl's theory are not obviously applicable because of over-idealization on Pearl's behalf (Morgan 2004) or because no clear method exists to translate between Pearl's theory and pre-existing statistical methods (Neuberg 2003; Hoover 2003).

Unfortunately a detailed account of these criticisms and analysis of Pearl's defence against them falls outside the aim of this thesis. As such I leave these issues aside and move on to consider criticisms of Pearl's regimentation arising in the philosophical literature. However, I emphasise to the reader that the partition introduced above, whilst not entirely artificial, is not entirely natural either. As such I direct the reader's attention back to section 2.1 where I discuss in some detail Pearl's attempt to draw a line between the statistical and the causal. On reflection, the discussion undertaken in that section might be seen to point towards several unresolved foundational issues in the field of statistics that are (arguably) philosophical in nature. Moreover, it might further be argued that some of the criticisms raised below by philosophers will be recognised as relatives of those criticisms offered by statisticians[177].

There are several approaches taken by philosophers who criticise, what is on my interpretation, Pearl's regimentation. Some are (or were at one time) sympathetic to Pearl's project and so tend to offer specific criticisms along with proposed solutions. Others deny the efficacy of the project either in part or in its entirety and so focus on exploiting perceived weaknesses in either the formalism or with the account's key

---

Pearl's semantics are too permissive. I do not discuss such issues directly.

[176] For discussion see Pearl (2002c, 2003a).

[177] One example of philosophical criticism offered by a statistician is the criticism offered by David Freedman. I recount one such criticism below. Furthermore, I place counterexamples to the formal component of Pearl's account into the statistical basket since such counterexamples typically require technical fixes. However, some of these counterexamples are wielded by philosophers with the aim of making deeper criticism of the account. For instance see Menzies (2002).

assumptions. In discussing the acceptability of Pearl's answer I will only focus on the criticisms from those of the latter persuasion. This group seek to challenge Pearl's answer on the basis that it does not even get off the ground because one or more of its conditions is either incoherent or lacks applicability even by its own lights. Call these criticisms the 'no-answer' criticisms. Each and every conjunct of Pearl's regimentation has been criticised in the literature[178]. I will consider several of the more prominent criticisms in a moment.

Since Pearl's answer is, I hold, the conjunction of conditions 1-5, Pearl's answer will be proven unsatisfactory in case any one condition is shown to fail. Even so, recall that Pearl is committed to the view that some causal claims remain viable without every conjunct being met. This is another way of saying that some of the criticisms targeted at each individual condition, granted they are successful, would be more damaging than others. For instance, Pearl's answer would face a significant set back were the Causal Markov Condition to be shown untenable. In light of this I will provide the greatest focus in the following discussion on conditions 2, 3, and 5 respectively and only superficially focus on the remaining conditions[179]. Although I hold that a defence against the criticisms set out below is straightforward I will not entertain any defence of Pearl's regimentation until after the criticisms have been levelled. I will then argue that each criticism is misplaced. Some discussion will be necessary to sketch out a more appropriate target. I suggest that this should be Pearl's objective account.

The criticism commonly levelled at the condition of determinism is straightforward. Condition 1 is, so the criticism goes, a non-starter because, as contemporary physics tells us, the world is fundamentally indeterministic. Hence, a condition that asserts the relation between causes and effects is deterministic in structure is either false or

[178] However, I can find no direct criticism of Pearl's employment of minimality. Minimality may be criticised indirectly however. I discuss this further below.
[179] But, having said that and, and as will become clear below, each criticism harbours further misgivings against Pearl's answer. Moreover, were those who criticise Pearl's answer provided ample time within which to articulate their case I am certain that what are now specifically focussed challenges would turn into challenges against more than one conjunct of Pearl's answer.

lacks applicability and the same goes for the answer to which it is part[180]. Nancy Cartwright puts this challenge in the following pragmatic way:

> [...] for most cases of causality we know about, we do not know how to fit even a probabilistic model, let alone a deterministic one. The assumption of determinism is generally either a piece of metaphysics that should not be allowed to affect our scientific method, or an insufficiently warranted generalisation from certain kinds of physics and engineering models (Cartwright 2000: 13).

There are several challenges levelled at condition 2 throughout the philosophy of science literature. I wish to consider one such challenge in particular but first cite some other prominent criticisms. A general challenge to the CMC is that it cannot be a universal condition for either one of two reasons. First, the CMC faces counterexamples. Second, there are cases of causation where it is far from obvious how the CMC is to be applied (Cartwright 2000: 16-17).

The counterexamples to the CMC appear in various forms and are often accompanied by counterexamples to Reichenbach's (1956) PCC and fork asymmetry theory of causal direction[181]. The relevant class of counterexamples may be traced back (at least) to van Fraassen (1980)[182]. The thrust of these counterexamples is that there exist causal forks (structures) where, contrary to the CMC the common cause represented by the structure fails to screen off the correlation between its effects. Typically, the examples involve causal forks where the common cause and its effects are microscopic phenomena. However, some cases are presented where microscopic common causes are imbedded in scenarios that involve macroscopic effects. The

---

[180] I am uneasy about the notion of 'deterministic structure' as the notion appears in Pearl. The unease arises from being unsure whether Pearl is making a descriptive claim about model structures or actual physical systems or both. I take a deeper look at this issue when considering Pearl's answer on the second tier.

[181] For example Horwich (1987).

[182] But see also the work of Salmon (1984: 168-178), Cartwright (1995, 1997), Hausman and Woodward (1999) and Martel (forthcoming).

relevant point this class of counterexamples illustrate is that the CMC is not sufficient, as it is presumed to be by its proponents, for the identification of all causal models. Alternatively, the point is sometimes dressed as a sceptical worry along the following lines. Because Pearl's account is bound by the CMC it runs the risk of misidentifying or ignoring altogether non-Markovian causal structures, where it is conceivable that such structures are in fact commonplace.

Cartwright frames another, related, criticism[183]. There surely are cases of causality where the CMC is applicable but, likewise, there are cases of causality where the CMC appears to fail. In the former instance the CMC is trivial and in the latter it is irrelevant. However, in between these two cases there are instances where it is unclear how one is to employ modelling techniques that presuppose the CMC. In such cases modellers are forced to make difficult judgements about whether or not, and in what way a specific causal relationship exists. One variable can contribute to the production or prevention of another variable in a great variety of ways. As a result reliable tests for whether one variable causes another must be finely tuned to how the cause functions to produce its effect almost case by case. In these instances it is far from obvious how the modeller is to employ the CMC and draw causal inferences (Cartwright 2000: 18-20). In summing up the problem Cartwright states:

> The term 'cause' is highly unspecific. It commits us to nothing about the kind of causality involved nor about how the causes work. Recognising this should make us more cautious about investing in the quest for universal methods for causal inference. Prima facie, it is more reasonable to expect different kinds of causes operating in different ways or imbedded in differently structured environments to be tested by different kinds of statistical tests (Cartwright 2000: 4).

[183] This criticism forms part of a broader attack on the Screening Off relation. However, Cartwright considers the latter condition to be a part of or at least a corollary to the CMC. See Cartwright (2000: 5) for discussion. Besides the stated criticisms, Cartwright offers several others. See for instance Cartwright (2003, 2002a).

The key challenge to condition 2 I wish to consider is due to Berkovitz (2002). The criticism comes in two parts. The first part of the challenge is general in that it refers to not just the CMC but also terms such as 'cause' 'influence' 'correlation' 'common cause' 'distinct event' and so on as these terms come to be recruited into the formal component of Pearl's account. The accusation is that the account takes several or all of these terms as primitives and imposes upon them formal constraints in the shape of axioms abstracted from and inspired by the intuition of those steeped in the practice of modelling and analysis. It then seeks to investigate the necessary consequences of these axioms and test their relevance against familiar problematic scenarios and competing methods of causal inference (Berkovitz 2002: 241-242). This approach to causality is problematic for the following reasons. The plausibility of the CMC depends on those principles upon which it is based, such as the PCC. But, the PCC is not really a principle so much as the blueprint or schema for a principle because the terms that form part of the principle's definition such as 'common cause' 'distinct event' 'correlation' and so forth may have as a matter of fact different specifications and different specifications yield different principles. Hence, the acceptability of condition 3 cannot be assessed until the exact meaning of terms such as 'causal', 'common cause,' 'correlation' and so on, are specified (Berkovitz 2002: 242). But since these meanings are not forthcoming (by the very nature of the project) Pearl's answer is unsatisfactory. The second part of the challenge questions the reliability of the causal inferences sanctioned by Pearl's axiomatic system. Since the system depends crucially on the availability of domain specific causal knowledge, which may be unobtainable when the terms 'causal', 'correlation', 'distinct event' and so forth are not sufficiently specified no inferences can be drawn on behalf of the system (Berkovitz 2002: 242)[184]. So, again, condition 2 is a non-starter in its present form and so Pearl's answer is unsatisfactory.

[184] A further point may be pressed against Pearl were his answer to this challenge to offer a univocal definition of the terms under question. Since, Pearl may be asked to explain why different disciplines in the (social) sciences appear to mean different things or indeed nothing specific at all by the term 'cause' and other terms under question. On this point see Woodward (2002b).

The substance of this last challenge relies (in part) on the fact that Pearl's use of the CMC is related to the PCC, which it is by Pearl's own admission (Pearl 2000a: 30, 58, 61). But the PCC, even granted its status as a principle rather than merely the schema of one, has more than one formulation. Which formulation Pearl is committed to is not altogether clear[185]. This matters since the status of the PCC as a metaphysical, epistemological, or empirical principle, and so the status given to the CMC by Pearl's theory, depends upon the specific version articulated, as Berkovitz's criticism points out. Moreover, justification of the PCC, with the aim of supporting the CMC, may in turn require taking a particular stance in regard to additional conditions the PCC implies[186]. Part of the problem then is that Pearl (2000a) does not contain an adequate specification of the PCC[187]. For instance, Pearl's (2000a) few references to Reichenbach's PCC are recorded in the form of slogans rather then detailed remarks or definitions. Pearl speaks of

> ...several familiar relationships between causation and association that are usually associated with Reichenbach's (1956) principle of common cause–for example, 'no correlation without causation,' 'causes screen of their effects,' 'no action at a distance' (Pearl 2000a: 61).

Other comments suggest that Pearl holds the principle to be an 'assumption' which states:

> ...if any two variables are dependent, then one is the cause of the other *or* there is a third variable causing both (Pearl 2000a: 30 original emphasis).

However, Reichenbach took the following statement to express the PCC:

---

[185] And so must be inferred, which I do below.
[186] I have in mind here the Cause-Correlation Link and the Screening Off relation as well as assumptions about the uniformity of nature.
[187] This challenge might easily be extended to several of the conditions making up Pearl's answer. See below for details.

> If an improbable coincidence has occurred, there must exist a common cause (Reichenbach 1956: 157),

by which Berkovitz thinks Reichenbach meant:

> Any representative correlation between distinct events or quantities, neither of which causes the other, must be due to a common cause. Where 'correlation' means positive statistical dependence and where 'representative' denotes the appropriate relative frequencies given the population within which the correlation apparently appears (Berkovitz 2002: 243)[188].

Note that the latter specification of the PCC does not include a screening off condition nor any explicit reference to contiguity requirements. It is arguable then that Pearl confuses the PCC with a distinct principle called the Cause Correlation Link (CCL)[189]:

(CCL) Any non-accidental correlation between two distinct events or quantities is due to either (i) a causal connection between them, or (ii) a common cause, or (iii) both (i) and (ii),

The CCL appears to be an adjunctive motivating principle of (statistical) inference or, at best, a corollary of the PCC (Berkovitz 2002: 241). Furthermore, Reichenbach held the PCC to pertain to probabilities interpreted as (relative) frequencies, not as degrees of belief as does Pearl, a fact that, for some, may compound the problem posed by the present criticism (Reichenbach 1976: 126). Exactly which specification of the PCC Pearl is committed to then remains obscure. Fortunately it is possible to clarify the status accorded to the PCC by examining Pearl's comments about the CMC. Pearl (2000a) contains a level of detail sufficient to specify the status of the CMC. It is

---

[188] See also Reichenbach (1956: 163) and Salmon (1984: 159-168). Note also that Reichenbach thought the principle was true because of thermodynamic properties. Hence, I take it that Reichenbach thought of the principle as empirical rather then epistemic.
[189] This formulation appears in Berkovitz (2001: 241).

clear that Pearl is committed to an epistemic specification of the CMC since Pearl uses this condition as a filter through which to acquire and organise data. It is reasonable to conclude that Pearl thinks of the PCC in the same way[190]. So, one portion of the criticism Berkovitz offers against the CMC has been met through clarification, the remainder is yet to be dealt with.

The next challenge to consider is levelled against the condition Pearl labels variously as 'autonomy' or 'modularity.' The reader will recall that autonomy is the assumption that each child-parent relationship (i.e. each function) in a causal model represents a distinct mechanism that can vary independently of other such mechanisms present in the model (Pearl 2000a: 63). One of the key reasons the modularity condition is considered to be a non-starter is that, if taken to be universal, it faces many counterexamples. So many, in fact, that its negation is a more plausible candidate for inclusion in an account of causation. The thrust of this challenge is, therefore, that there is no good reason to support the assumption that causal processes are (in general if not universally) autonomous. That is, modularity is false as a matter of empirical fact. For instance, critics offer the following counterexample: autonomy may fail due to the fact that what at first appear to be intuitively distinct mechanisms turn out to be stand in the same spatiotemporal location in such a way that intervening to alter one mechanism without disrupting the other is impossible (Woodward 2002). Cartwright (2001) illustrates the point with a stylised scenario involving the mechanisms within a bread toaster:

> The expansion of the toaster's sensor due to the heat produces a contact between the trip plate and the sensor. This completes the circuit, allowing the solenoid to attract the catch, which releases the lever. The lever moves forward and pushes the toast rack open. I would say that the movement of the lever causes the movement of the rack. It also causes a break in the circuit.

[190] This interpretation gains further support from Pearl's comments on temporal asymmetry and causal directionality. Nonetheless it remains a fact that the details are found to be wanting. I note in passing that Pearl's view of the CMC and the PCC as epistemic principles or conditions makes Pearl's account

> Where then is the special cause that affects only the movement of the rack? Indeed, where is there space for it? The rack is bolted to the lever. The rack must move exactly as the lever dictates. So long as the toaster stays intact and functions as it is supposed to, the movement of the rack must be fixed by the movement of the lever to which it is bolted (Cartwright 2001: 72).

The upshot is that with this particular type of toaster design one cannot claim that the rack operates according to a different mechanism than that of the circuit despite the fact that they are intuitively distinct (Woodward 2002). According to Cartwright such instances, far from being novel, are in fact common features of the world.

Cartwright raises a related objection to condition 5–the stability, or, faithfulness condition. The thrust of Cartwright's criticism of stability is that the condition is commonly violated in scientific practice and, more to the point, violating stability is often a scientific goal in itself. For instance, Cartwright points out, the stability condition is violated whenever two causal processes are equally effective and cancel each other out. Moreover, Cartwright assert that, far from being pathological, violating stability is one of the ways that scientists and policy makers "minimise damage in our social systems and in our medical regimens" and allows technicians and others to construct many useful technological devices (Cartwright 2000: 16, 17). The lesson Cartwright draws from the criticism is that, as with the Markov condition, the stability condition is an unreliable measure of causal processes, and so, any conclusions drawn about causal relations based upon methods that presuppose stability can only be as secure as the prior understanding one has of the structure under investigation (Cartwright 2000: 17). This is not the only criticism Cartwright has of stability. Cartwright also criticises the stability condition on the grounds that the condition presupposes causal structures are fundamental and fixed whereas (probabilistic) parameter values are free to vary. Cartwright claims that this view is surely incorrect in most instances. For, says Cartwright, it is more often the case, if

---

distinct from that of SGS who I gather take the principles to range over frequencies of variables from well-defined classes and thus to be primarily empirical.

not always a general feature of causal systems, that structure and parameters arrive together. That is, contrary to the stability condition, probabilities and causal structures constrain each other and if the probabilities are fixed then certain causal structures are ruled out (Cartwright 2003: 262-263).

Detailed criticisms of Pearl's minimality condition are rare. Indeed Pearl notes "... few have challenged the principle of minimality (to do so would amount to challenging scientific induction)" (Pearl 2000a: 61). It is, however, not difficult to construct a criticism of minimality. First, recall that the principle of minimality relates to causal structures[191]. The principle is invoked to guide decisions concerning model preference in response to problems related to observational equivalence discussed in section 2.0.1:

> In principle, since V is unknown, there is an unbounded number of models that would fit a given distribution, each invoking a different set of 'hidden' variables and each connecting the observed variables through different causal relationships. Therefore, with no restriction on the type of models considered, the scientist is unable to make any meaningful assertions about the structure underlying the phenomena. [...] Likewise, assuming [the set of variables V is equivalent to the set of observed variables] but lacking temporal information, the scientist can never rule out the possibility that the underlying structure is a complete, acyclic, and arbitrarily ordered graph–a structure that can *mimic* the behaviour of any model, regardless of the variable ordering (Pearl 2000a: 45 original emphasis).

Pearl (2000a) provides the following statement of the principle of model minimality:

> [...] following standard norms of scientific induction, it is reasonable to rule out any [structure] for which we find a simpler, less elaborate [structure] that

[191] More correctly, minimality relates to latent structures–causal structures with unmeasured variables.

> is equally consistent with the data. [Structures] that survive this selection process are called *minimal* (Pearl 2000a: 45 original emphasis)[192].

Based on these points it is clear the condition involves a decision procedure that encodes for a form of abductive inference. Since minimality is reminiscent of the so-called 'Inference to the Best Explanation,' one may dub minimality 'Inference to the Best Structure' or IBS for short. I take it that IBS is a part of what Pearl means by 'scientific induction.' I further note that Pearl's observation that few seek to challenge IBS must be tempered by the fact that no one can claim to have adequately vindicated it either. It is at this point important to be clear about what exactly IBS is.

IBS is a procedure for locating the simplest model structure–that is consistent with a specific set of observations–from a class of model structures. Pearl provides the following characterisation of 'simplest model structure.' The simplest model structure is called the 'preferred' model structure. Call a model structure a latent structure just in case:

> The model structure is a pair $L = \langle D,O \rangle$ where D is a causal structure over V and where $O \subseteq V$ is a set of observed variables[193].

One latent structure $L = \langle D,O \rangle$ is strictly preferred over another latent structure $L' = \langle D',O \rangle$ (written $L \preceq L'$) iff:

> D' can mimic D over O–that is, iff for every Θ_D there exists a $\Theta'_{D'}$, such that $P_{[O]}(\langle D',\Theta'_{D'} \rangle) = P_{[O]}(\langle D,\Theta_D \rangle)$. Two latent structures are equivalent (written $L' \equiv L$) iff $L \preceq L'$ and $L \succeq L'$[194].

---

[192] I have replaced the word 'theory' with 'structure' throughout. I find the latter term more appropriate in the context. Note that Pearl (2000a: 45) equates a causal structure with a scientific theory but does not elaborate on the point. The formal details are available in Pearl (2000a: 45-46).
[193] This definition is taken from Pearl (2000a: 45).
[194] This definition is taken from Pearl (2000a: 45-46).

Hence, a latent structure L is minimal with respect to a class of latent structures **L** iff:

> No member of **L** is strictly preferred to L–that is, iff for every $L' \in \mathbf{L}$ we have $L \equiv L'$ whenever $L' \preceq L$[195].

Minimality becomes a procedure when model structures are compared and accepted or rejected on the basis of these definitions. The question naturally arises as to why minimality is an adequate basis on which believe in one structure over another if it is the case that any putative causal model structure may be mimicked by a latent structure containing alternative causal relationships.

There are apparently several good reasons against doing so[196]. The nub of IBS is that, given the data, the scientist can utilise IBS to arrive at the simplest DAG consistent with the data. At least two steps are involved in this inference. First, there is a weighing of evidence in relation to distinct structures from the class of latent structures **L**[197]. This step arrives at an equivalence class of minimal structures. The second step involves claiming that the structure(s) identified by the first step are more likely to be the correct structure of the processes that generated the data than not. The second step arguably depends upon the prior belief that the true structure is more likely to be found amongst the structures of the equivalence class than not. It is straightforward to construct a criticism of minimality by denying it is rational to hold this prior belief. Denying the rationality of the prior belief requires attacking attempts to justify it. Since there are several possible justifications there are likewise several avenues of attack. The most relevant justification in the present context involves the attempt to vindicate the prior belief by arguing that humans are by nature predisposed to hit upon the right range of structures. This justification is relevant because it underpins Pearl's claim that human capacity for causal knowledge is the result of a process of natural selection (Pearl 2000a: 26 n12, 244). But, so the criticism runs,

---

[195] This definition is taken from Pearl (2000a: 46).

[196] The following criticism is an adaptation of a criticism of abduction offered by van Fraassen. For the details of the original see van Fraassen (1989: 142-144).

such a justification is illegitimate because it requires that evolution select for predispositions towards correct judgements, which amounts to the claim that there is such a thing as selection pressure on future contingents. The upshot of the criticism is that, however natural our inclination may be toward a class of preferred structures the inclination itself cannot be relevant information about the correctness of the structures in the class.

Now that I have considered the apparent shortcomings of Pearl's answer, I can proceed to discuss possible responses Pearl may utilise in defence. As I mentioned above, I believe each of the criticisms may be defeated with the one strategy. Pearl has the option, given the two-part distinction, to show that each challenge is misplaced. The general idea behind the defence is to first show that each of the challenges (in effect) targets what amounts to the theory's first part and then demonstrate that each criticism fails to hit that target. This strategy, of course, stands or falls on the cogency of the two-part distinction.

The misplacement defence begins with the following observation. None of the criticisms draws nor obeys any explicit distinction between providing an account of causal modelling on the one hand and devising an account of causation on the other. As a result none of the criticisms takes into consideration the disparate goals of the respective accounts. Instead, each of the criticisms I have considered presupposes that Pearl's theory consists of an explication of causation as it is in the world, or sets out an account of causal inference that somewhere involves the identification of bona fide *causal* connections[198]. For instance, note that each of the criticisms presupposes that Pearl's theory is a representative of a distinct school of thought or approach to the analysis of causation. Berkovitz addresses his remarks towards the (contemporary) literature on causal inference (Berkovitz 2002: 238-239), Cartwright levels her criticisms at the 'Bayes Nets approach' to causality (Cartwright 2003: 253-254) and

[197] I have not discussed this process in great detail since I will not offer any criticism of it. Pearl (2000a) discusses the process in various contexts on pp 45-46, 51-55, 61-64, 274-275.
[198] Berkovitz (2002), Humphreys and Freedman (1996) are representatives of the former view and Cartwright is a representative of the latter.

other authors generally agree that Pearl belongs to a category of theories that attempt to give account of causality based on (graphical/statistical) modelling methods and principles[199]. This oversight, I believe, leads these authors to misplace their respective criticisms. Furthermore, it is no coincidence that each criticism is levelled at one or another conjunct making up Pearl's view of what puts the 'causal' into 'causal Bayesian network'. After all, this does appear to be Pearl's primary focus. Even so, my point is that the criticisms are misplaced precisely because each conjunct only represents the regimentation's answer, which is the part concerned to articulate an account of causal modelling, not an account of causation *per se*. And, in as much as each criticism aims to uncover the failure of Pearl's account of causal modelling to specify the nature of or to identify causal processes, the criticisms miss their mark. In short, each criticism is misplaced since each criticises a condition of the regimentation for failing to take proper account of a subject matter only appropriate to the objective account of Pearl's theory[200].

I now detail the defence of each condition taken in turn, beginning with condition 1. According to condition 1, the values of effect variables, as they appear in a model, are completely determined by the values of the effect variables direct cause (and error) variables as they appear in the model. But so what? That the independent variables determine the value of the dependent variable is a condition that allows the model to be useful and informative. Very little would be achieved by banning a modeller from building models sufficient for prediction. Pearl explicitly points out that the relevant notion of determinism is subjective and that deterministic functions offer an adequate representation of that notion. It is left to the modeller to decide at what level of abstraction causal statements are meaningful. When the modeller settles on the appropriate level the modeller may then proceed to evaluate the truth of those causal

---

[199] Wolfgang Spohn's (2001) expresses this view explicitly. Other research that attracts philosophical attention as a representative of this school-of-thought, besides Pearl (2000a), is SGS. See also the critical review of Pearl (2000a) offered by Gillies (2001) and Hitchcock (2001). Both of these authors read Pearl to be offering an analysis of causation. Furthermore, the assertion that Pearl's account of causality, as is the case with most agency accounts of causation, is circular and hence not an analysis of causation after all, is precisely to view the account as a failed analysis.

[200] Korb and Wallace (1997: 550-551) make a similar point regarding the use of linear modelling methods to draw causal inferences from measurements of causal processes that are not linear.

statements. Banning determinism prevents the modeller from finding the appropriate level of abstraction and so prevents model construction altogether. Deterministic functions offer a far more simple approximation of the complexity of worldly causal processes than do stochastic functions. It is not important whether or not the natural world is truly deterministic at the level causal models operate[201].

The criticisms of condition 2 are misplaced in as much as the criticisms presupposes the CMC plays the role of an empirical assumption in Pearl's account of causal modelling. The presupposition is that actual physical processes confirm a specific category of independence relationships. But, as Pearl (2000a) is at pains to point out, the CMC

> [...] is more a convention than an assumption, for it merely defines the granularity of the models we wish to consider as candidates before we begin the search. [...] After all, investigators are free to decide what level of abstraction is useful for a given purpose, and Markovian models have been selected as targets of pursuit because of their usefulness in both prediction and decision making (Pearl 2000a: 44, 61).

As such, Pearl freely admits his willingness to ignore non-Markovian causal structures. In effect Pearl is asserting that models built to conform to the CMC prove to be predictively and/or explanatorily successful, and therefore useful, in a broad range of circumstances[202]. Enshrining the CMC as part of a formalism built to aid the task of causal discovery is to enshrine prediction and explanation as the goals of causal modelling[203]. Ironically, Cartwright makes this point on Pearl's behalf whilst castigating Pearl (2000a) for not taking note of causal diversity and not including a

---

[201] Again, the distinction between how and why is involved. Also note that Pearl (2000a) offers an extension of his modelling methods to cases involving objective chance on p 220. Whether or not the extension is correct the attempt demonstrates Pearl's willingness to find a work around for difficult cases without dispensing with deterministic functions.

[202] Though in fact, the successes may be few. See the discussion of the identification problem below.

[203] Of course, that there is a conceptual component to Pearl's theory requires that these terms are also provided with their meaning from conditions such as the CMC. Models that adhere to the CMC are explanatory in part because the CMC adds to the definition of causality and explanation in science.

contiguity requirement. Cartwright follows Shafer (1997), who points out the pitfalls involved in attempting to specify causal relationships:

> Experience teaches us that regularity can dissolve into irregularity when we insist on making our questions too precise, and this lesson applies in particular when the desired precision concerns the timing of cause and effect [for instance] (Shafer 1997: 6).

Based on Shafer's point, Cartwright asserts that natural causal relationships are extremely complex and fragile:

> [Shafer's] point is reinforced when we realise that the kind of physical and institutional structures that guarantee the capacity of a cause to bring about its effect may be totally different from those that guarantee that the causal message is transmitted. Here is an example. My signing a check at one time and place drawn on the Royal Bank of Scotland in your favour can cause you to be given cash by your bank at some different place and time, and events like the first do regularly produce events like the second. There is a lawlike regular association and that association is a consequence of a causal capacity generated by an elaborate banking and legal code. Of course the association could not obtain if it were not regularly possible to get the cheque from one place to the other. But it is; and the ways to do so are indefinitely varied and the existence of each of them depends on quite different and possibly unconnected institutional and physical systems: post offices, bus companies, trains, public streets to walk along, legal systems that guarantee the right to enter the neighbourhood where the bank is situated, and so on and so on (Cartwright 2000: 15-16).

Hence, a putative regularity may actually involve many and varied causal relationships as producers and sustainers. But Pearl presupposes just this fact when he asserts that the decision of the appropriate level of abstraction at which point the

useful properties of prediction and explanation are lost remains in the hands of the investigator. But this just means that ascertaining whether or not and in what way the structure of a model corresponds to underlying physical processes in the world is not the goal of modelling. In other words, the latter task stands to the side of the regimentation[204]. So, basing a challenge to condition 3 on such grounds is misguided. So too is pointing out that in many cases it is not obvious how to apply the CMC to a specific scenario. The onus of constructing an appropriate model for a given problem rests with the individual(s) performing the investigation and not with the formalism chosen as an aid to that end[205]. A similar point holds for the criticism that many key terms in Pearl's account remain undefined. This criticism ignores the fact that there is, in practice, interplay between the semantics of the formalism on the one side and the content and context in which models that employ those semantics are constructed on the other. Expecting a (largely formal) modelling procedure to furnish the meaning of 'cause,' 'effect,' and their cognates and criticising it when it does not is akin to criticising a decision procedure after finding one does not thoroughly understand the meaning of the propositions of an argument the decision procedure has deemed valid. The task of the modelling procedure is to aid in the identification and organization of elements within the domain in which the data are gathered that enable successful prediction and explanation. The specification provided to the term 'cause' within Pearl's account is sufficient for this task[206], and the further details expected by this challenge can only be furnished through specific application and mental exercise on behalf of those carrying out an investigation. Therefore, it is to specific applications of the formalism that criticism of a lack of meaning be attached and surely not to the formalism itself.

The key criticism of the autonomy condition was to the effect that it fails to hold universally. This claim is false. A model containing a causal relationship that does not display the possibility of alteration without affecting other causal relationships in the model would indeed be destructive to Pearl's account. But, far from being a

[204] Indeed, these questions fall in the domain of the second tier.
[205] Though the formalism is built to be useful.

counterexample, on the semantics of Pearl's account such a model is simply not well-formed. All well-formed models conform to the autonomy condition. What then is the point of attempting to construct counterexamples to autonomy? To my mind what the toaster (and other similar) examples illustrate is nothing more than the importance of building models appropriate to one's purposes. That is, one should strive to build models capable of answering all the causal queries one might care to answer about the process under investigation. If this is the thrust of the challenge then it is good advice, but the advice hardly constitutes a counterexample to Pearl's account. Perhaps then this is not the thrust of the challenge and I have built a straw-person out of it. Instead, it may be the aim of this challenge to attack the autonomy condition for failing to correctly describe the mechanisms that (at least intuitively) form part of the system under study. But, this being the case, a similar defence seems appropriate; what counts as the correct description turns on the interests and context of the investigator who constructs the model–their interests determine the granularity of the model and what the model describes follows suit[207].

A similar defence is also appropriate for defending against the criticisms levelled at the stability condition. The universality of the stability condition is another defining characteristic of the account's semantics and it cannot be altered without thereby altering the entire nature of Pearl's theory. Whether or not stability is violated in specific instances is again a matter for the modeller and only to a minor extent the modelling procedure. Similarly, the fact that the stability condition's holding of a model is a reflection of the degree of prior understanding the modeller possesses of the system they have under investigation, does not count against Pearl's theory. A modeller's knowledge of what they are modelling is in many cases very good. Good enough that, with the aid of a formal modelling procedure, the modeller can provide excellent explanations and predictions. But, often the modeller's knowledge of the system under investigation is poor. One doesn't necessarily blame the procedure for failures in such cases. Moreover, there are numerous instances where a modeller may

[206] Setting aside faults in the formalism itself.

be investigating a causal system but be content to draw inferences about correlations and ignore causal information altogether. For instance, an investigator may discover that inferences based upon correlations guide predictions, whereas, with the addition of causal information, predictive power is lost[208]. Ignoring stability to reach such ends does not count against the formalism in any sense other than to delimit its applicability. Last, is it really the case that structure and parameters are fixed and arrive together? Not as the matter concerns the construction of a model. Model construction may occur in many and varied ways but typically begins with parameters and ends with (the reasonable acceptance of a) structure. Surely, it would be better to say that whether specific or proposed parameters and structure are fixed together is to be discovered case by case[209]. The criticism has a metaphysical dimension to it. Since the actual world is just the succession of particular matters of fact, so the argument goes, it is a general feature of bona fide causal processes that structure and parameters are invariant in relation to one another. What to say here? Even if this were true it is a discussion pertaining to the objective account and not to the regimentation. And, again, rather than being an assumption of the regimentation, it is just the sort of thing it is designed to model. For Pearl, invariant parameters are invariant because they track model structures.

In a similar vein defending the minimality condition will lead me to consider the justification of modelling assumptions in the following section. The criticism of minimality I considered involved attacking the rationality of basing one's beliefs about the correct causal structure on the fitness of our natural inclinations. The weakness of the criticism is its supposition that Pearl's account of causal modelling requires that the investigator not merely assent to the choice of structure but actually believe it to be correct. Since Pearl's account requires no such thing employing

---

[207] However, this criticism treads close to another–the failure to address the identification problem. See below for details.

[208] The example in mind involves data complexity. In cases where there are several levels of complexity present within a causal system, the level at which predictive power is greatest may not withstand a causal interpretation. This occurs, for example, in marine ecology modelling. It remains to be seen whether Pearl's theory can be extended to such cases.

[209] Which is in effect the thrust of the stability condition.

minimality does not commit one to the prior belief that the true structure is more likely to be found amongst the structures of the equivalence class than not. Indeed, minimality is not the only condition that pertains to structure preference; being only one of several conjuncts to Pearl's account of causal modelling. Hence, despite the comments Pearl apparently makes to the contrary, the minimality principle does not force epistemic assent and so nothing of consequence is lost if the true structure is not located. This fact does, however, propel the issue of justifying causal models to the fore.

3.4 Interpreting and Justifying Causal Models

The discussion of Pearl's regimentation has returned me to the subject matter of Pearl's objective account. Recall that the mistaken presupposition common to each criticism was that Pearl's theory articulates an analysis of causation solely in terms of what amounted to the five conjuncts of the regimentation[210]. I have claimed that this defence is well supported by Pearl (2000a) despite the fact that Pearl does not explicitly articulate his account in the terms of the two-part division I set out. Whatever the level of support, the guiding idea was to defend Pearl (2000a) from criticism that failed to take into consideration the disparate goals each part plays a role in achieving and, in so doing, increase the depth of the two-part interpretation. In support of this line of defence Pearl (2003b) submits the following challenge:

> ...these confusions and difficulties stem primarily from a reluctance to communicate causal questions in a formal language such as the one offered by the atomic framework [set out in Pearl (2000a)]. In my experience, I have found that, invariably, questions about interventions and experimentation, ideal as well as non-ideal, practical as well as epistemological, calculational as well as interpretive can be formulated precisely and managed systematically using the atomic intervention as a primitive notion. The same applies to questions about the 'correctness' of causal models. I will thus end [...] with a conjecture (or a challenge) that any philosophical question, disagreement, or difficulty concerning causal objects or causal relationships can be resolved if expressed formally in the language of atomic interventions and reduced to a mathematical exercise in the calculus of $[P(y \mid do(x))]$ (Pearl 2003b: 4).

[210] This is not the only point about which the criticisms are mistaken. See below for details.

The challenge is not as outlandish as it first appears[211]. I interpret Pearl to be asserting that difficulties and disagreements–especially those concerning alleged counterexamples–had by philosophers and others over the nature of Pearl's theory should be (whenever possible) formalised and a solution attempted before negative conclusions are attributed to the account itself. The distinction I have drawn between Pearl's regimentation and Pearl's objective account helps explain the point. Pearl does not accept, I argue, that the criticisms and alleged counterexamples I have discussed above are appropriate if levelled at the various components of the regimentation and formalism of the theory's first tier. Instead, Pearl appears to be suggesting the criticisms operate at the level of and presuppose specific models (or structures) of causal scenarios. But, of course, at the level of the model and where a specific causal scenario is under investigation, the problems alluded to by each challenge are commonplace and it is through utilising the components of the regimentation that Pearl thinks such problems may be at least clearly stated if not rendered tractable, a point that I will return to momentarily.

There are consequences for adopting this strategy. One consequence of wielding the sort of defence I have on Pearl's behalf is that the theory does not escape criticism completely unscathed. If I am correct, and the criticisms I consider are indeed misplaced, then it is natural to ask whether they might not fare better were they to be reformulated and aimed at another component of Pearl's theory. After all, the challenges have not been shown to be incoherent and so it remains a live option to refocus or reformulate each of them[212]. I suggest the most appropriate reformulation

---

[211] Indeed, if one took the scope of 'any philosophical question, [...]' widely enough then Pearl's challenge is easily met. For instance Pearl says nothing about adicity and his identification of variables with events (represented as propositions) sidesteps the issue of whether causal relata are immanent or transcendent. Furthermore, given the two-tier interpretation of Pearl (2000a) these types of metaphysical questions would appear to belong (if anywhere) to the second tier, which of course Pearl says very little about. I take it then that such questions do not have ready answers even if they could be reduced to mathematical exercises in the do-calculus.

[212] Refocussing the challenges onto Pearl's objective account is I think the best option. But, since I have been unable to uncover the key characteristics of the objective account, I cannot go far towards assessing the success of otherwise of such a new round of challenges. In any case, such challenges are all too common within the domain that is the subject matter of the second tier. Suffice it to say that, in this domain, questions of universality, direction, determinism and so forth are faced by everyone who takes causal analysis seriously and not just Pearl.

would see the challenges levelled at Pearl's objective account. But, having said that, and being mindful of Pearl's comments cited in relation to the matter, I shall not now attempt a reformulation nor detail a new round of criticism of this portion of Pearl's theory. It is my view that, once clarified, each challenge will be either less urgent than first appearances suggested–since almost every modelling procedure faces the sort of problems these challenges are intent on recording–or merely a corollary to another pre-existing problem of how to justify modelling assumptions–which, again, is a problem every account faces. Hence, in lieu of any attempt to repair and refocus existing criticism I will instead examine the role modelling assumptions play in Pearl's account. The examination boils down to articulating what, on Pearl's theory, a causal model is a model of.

Recall that on Pearl's theory models have two primary roles. One role is to aid the investigator's reasoning about those subject matters in the sciences that inevitably involve causality. The other role is to act as a testable representation device for causal relationships in the context of a specific study design. These roles immediately involve issues of 'identifiability', a notion that will take a moment to outline.

Put simply, a model is said to be identifiable when it contains a sufficient number of independence assumptions to permit the quantification of the effect one set of modelled variables have on other modelled variables (Pearl 2000a: 91). Identifiability can become a problem when, in cases where any one of several (distinct) models can generate an identical distribution, the desired quantity the investigator seeks to quantify might not be discernable unambiguously from the data. Pearl (2000a) sees this difficulty as an instance of the general problem of how to estimate a causal quantity from passive observations (Pearl 2000a: 77). According to Pearl (2000a), having a solution to identification is

> ...essential for integrating statistical data with incomplete causal knowledge of $\{f_i\}$, as it enables [the investigator] to estimate quantities Q consistently from large samples of [probability distributions] P without specifying the

details of M; the general characteristics of the class $\boldsymbol{M}$ suffice[213] (Pearl 2000a: 77).

The general approach taken towards solving identification problems is to specify a condition for model 'identifiability' and then assess a model's quantities against such a standard before moving on to estimation and model testing in case the assessment of identifiable quantities has been positive. In statistics generally, identifiability is the property a statistical model exhibits such that it may be estimated consistently from a suitably large amount of data on the system the model represents (Dodge 2003:192). In fact, a statistical model may be 'under-identified', 'just-identified', or 'over-identified'. Shipley (2000) provides the following simple example to illustrate each category. Given the equation $y = 2x + z$, together with the information that $x = 1$, it follows that more than one combination of values for the variables y and z will solve the equation. In cases of this variety the equation is said to be under-identified. If instead one knows both that $x = 1$ and $z = 3$, then it follows that the variable y can take only one value; 5, and so the equation is called just-identified. However, if it is known that $x = 1$ but that two distinct estimates of z are 2.5 and 3.5, then the equation is over-identified (Shipley 2000: 145). Of the three possibilities it is under-identified models that investigators seek to avoid.

At one time it was common wisdom that identifiability attaches to models taken in their entirety. It is now common to speak of identifiability in terms of parameter identifiability. Hence, it is thought that to estimate a model it is first necessary to establish whether the model's (unknown or free) parameters are identifiable. If each of these (unknown) parameters is identifiable, then the model is considered identifiable. A similar line holds for causal models, but where, according to Pearl, since the aim of the causal modeller is to predict a *post-intervention* distribution from a *pre-intervention* distribution, identifiability amounts to assessing whether a model will permit the estimation of the causal effect of one variable upon another. (Pearl

[213] Where M is a causal model and $\boldsymbol{M}$ a class of such models. Granted a model M with variables X and Y an example of a estimable quantity Q is the causal effect of X on Y.

2003a: 297-298). Koopmans (1985) explains that the difference between identifiability in causal models, in contrast with identifiability of associational models, hinges on the requirement of causal models:

> [...] to predict the values of one or more [...] variables either under changes in structure that come about either independently of [the investigator] or under hypothetical changes in structural parameters [...]. [...] In such cases, the 'new' distribution of the variables on the basis of which predictions are to be constructed can only be derived from the 'old' distribution prevailing before the structural change, if the known structural change can be applied to identifiable structural parameters, that is, parameters of which knowledge is implied in a knowledge of the 'old' distribution combined with the *a priori* considerations that have entered into the model (Koopmans 1985: 122).

Pearl (2000a) offers the following specification of identifiability for causal models[214]:

> Let $Q(M)$ be any computable quantity of a model M. We say that Q is identifiable in a class $\boldsymbol{M}$ of models if, for any pairs of models M_1 and M_2 from $\boldsymbol{M}$, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$. If our observations are limited and permit only a partial F_M of features (of $P_M(v)$) to be estimated, we define Q to be identifiable from F_M if $Q(M_1) = Q(M_2)$ whenever $F_{M_1} = F_{M_2}$[215].

Pearl considers quantities (and so some models) not computable in this way to be 'non-identifiable'. In practice a quantity is non-identifiable when it is under-identified[216]. According to Pearl (2000a), models that are non-identifying limit the ability of the modeller to determine the quantitative relationship between a model's

---

[214] The following definition of identifiability is from Pearl (2000a: 77).

[215] This specification holds for estimating models where each parameter of the model is measurable. Semi-Markovian models and models used to estimate counterfactual quantities are given a slightly different treatment which is not important for present purposes.

[216] Or when identifiability criteria are compromised by data-related problems. In the case of the latter it may be possible to specify bounds on estimates of causal quantities. See chapter 8 of Pearl (2000a) and Pearl (2003a: 308) for discussion.

variables. For instance, it may not be possible to specify the direct effect one variable has on another in a non-identifiable model. In turn, non-identifying models provide no support for (and may even cast doubt upon) model structure. In the converse case, identifiable models can withstand further testing in terms of further observation or experiment.

Along these lines, Pearl (2004) acknowledges that in many cases an investigator is only interested in understanding the effect one variable has on another or wishes to corroborate only a small set of claims a given model implies about such relationships whilst ignoring other relationships implied by a given model as irrelevant. It is in instances such as these, Pearl finds, the investigator needs to know whether or not a relationship to be estimated is non-identifiable just as a consequence of specific assumptions the investigator has built into the model[217]. Similarly, in cases where the specific relationship of interest is over-identified, the investigator requires reassurance that when the relationship is estimated the result is not solely a reflection of the investigator's assumptions embedded in the structure of the model[218] (Pearl 2004: 2). In other words, the investigator's general concern regarding identifiability is to delineate relevant sets of assumptions from irrelevant sets in order to assess which assumptions are sufficient for uniquely substantiating a given claim implied by a model. In turn, this concern is underpinned by the investigator's interest in locating *bona fide* over-identified quantities, since these present an option of testability missing from just-identified quantities[219] (Pearl 2004: 1-2). Put simply, the investigator wishes to know whether the assumptions embodied by a model are sufficient for uniquely substantiating a given set of causal claims in order to be in a position to assess to what degree such claims are merely a reflection of those assumptions.

---

[217]The assumptions are typically related to the structural specification of the model and are typically a consequence of a priori considerations. See below for examples.

[218] Pearl (2004) labels this 'irrelevant over-identification'.

[219] Examples of model assumptions include the absence of directed paths or of variables in an equation, fixed coefficients in some equations, equality constraints between some parameters and so forth. See Pearl (2004) and Shipley (2000) for further discussion of modelling assumptions.

The question of what a causal model is a model of can thus be broken into two parts. Broadly, one part involves interpreting what a causal model says and the other part involves justifying the interpretation to hand. Hence, for a model to (meaningfully) represent causal relationships, its (relevant) parameters must be identifiable and its supporting assumptions must withstand scrutiny. Recall that Pearl identifies his theory with a collection of procedures that

> [...] facilitate the drawing of quantitative causal inferences from a combination of qualitative causal assumptions (encoded in the diagram) and nonexperimental observations (Pearl 2000a: 94).

Drawing such inferences amounts to providing a model with an interpretation. The interpretation is conditioned on a set of assumptions about the model's structural features. In turn, the interpretation stands or falls pending the quality of its justification. Pearl (2000a) addresses the matter in the following way:

> [...] causal assumptions in themselves cannot generally be tested in nonexperimental studies, unless they impose constraints on the observed distributions. The most common type of constraints appears in the form of conditional independencies, as communicated through the d-separation conditions in the diagram. Another type of constraints takes the form of numerical inequalities. [...] For example, we show that the assumptions associated with instrumental variables are subject to falsification tests in the form of inequalities of conditional probabilities. Still, such constraints permit the testing of merely a small fraction of the causal assumptions embodied in the diagrams; *the bulk of these assumptions must be substantiated from domain knowledge as obtained from either theoretical considerations or related experimental studies* (Pearl 2000a: 94-95 my emphasis).

Freedman (2002) agrees:

> The issue boils down to this. [Given a three variable model, for example,] does the conditional distribution of Y given X represent mere association, or does it represent the distribution Y would have had if we had intervened and set the values of X? There is a similar question for the distribution of Z given X and Y. These questions cannot be answered just by fitting the equations and doing data analysis on X, Y, and Z; additional information is needed. From this perspective, the equations are 'structural' if the conditional distributions inferred from the equations tell us the likely impact of interventions, thereby allowing a causal rather than an associational interpretation. The take-home message will be clear: you cannot infer a causal relationship from a data set by running regressions – unless there is substantial prior knowledge about the mechanisms that generated the data (Freedman 2002: 6).

So it seems only proper to question the role that theoretical knowledge plays in justifying modelling assumptions. Freedman (2002) does just this when he draws attention to the friction created by the attempt to draw causal inferences from a model where interventions are only hypothetical:

> We want to use regression to draw causal inferences from nonexperimental data. To do that, we need to know that certain parameters and certain distributions would remain invariant if we intervene. That invariance can seldom if ever be demonstrated by intervention. What then is the source of the knowledge? [...] 'Theory' seems like a natural answer, but an incomplete one. Theory has to be anchored in reality. Sooner or later, invariance needs empirical demonstration, which is easier said than done (Freedman 2002: 8).

Recall that Pearl intends that the appropriate (class of) causal models can demonstrably track the physical constraints present in the system under investigation just in case the causal inferences drawn on behalf of a model are valid. The forgoing discussion has highlighted the fact that to carry out the demonstration one must either

justify or discharge the assumptions underlying the model. It is precisely this attempt to 'anchor theory in reality' that leads to the difficult task of model justification:

> Given that the arrows and kernels represent causation, while variables are independent and identically distributed, we can use Pearl's framework to determine from the diagram which effects are estimable. This is a step forward. However, we cannot use the framework to answer the more basic question: Does the diagram represent the causal structure? (Freedman 2002: 15).

Freedman's scepticism implies that because models constructed according to Pearl's theory are vulnerable to problems of justification the models do not succeed in modelling causal relationships. In other words, Pearl's theory does not produce models that represent causality since the theory does not provide the means by which the models of the regimentation may be identified with the objective constraints mentioned by the objective account. To assess whether Freedman's point is sound consider the following example from Pearl (2000a).

Figure 1: Smoking and Cancer

As Freedman points out, the inference to cause in this model is conditional on (at least) three assumptions. First, genotype has no direct effect on tar deposits. Second, smoking has no direct effect on lung cancer. Third, tar deposits can be measured with reasonable accuracy. To highlight the difficulty of vetting modelling assumptions

Freedman asserts that none of these assumptions has been demonstrated by empirical studies:

> [Against the first assumption,] the lung has a mechanism–'the mucociliary escalator'–for eliminating foreign matter, including tar. This mechanism seems to be under genetic control. (Of course, clearance mechanisms can be overwhelmed by smoking). The forbidden arrow from genotype to tar deposits may have a more solid empirical basis than the permitted arrows from genotype to smoking and lung cancer. [The second assumption] is just that–an assumption. And [the third assumption] is clearly wrong (Freedman 2002: 14).

In other words, it is important to realise that identification problems are not solved just because a model meets an identifiability criterion. Indeed, the aim of making model identifiability decidable plays only a minor role in representing causal quantities[220]. Pearl acknowledges this:

> [...] the primary use of [identifiability methods] lies not in testing causal assumptions but in providing an effective language for making those assumptions precise and explicit. Assumptions can thereby be isolated for deliberation or experimentation and then (once validated) be integrated with statistical data to yield quantitative estimates of causal effects (Pearl 2000a: 95).

Pearl might therefore question Freedman's claim that the model of smoking and cancer receives little or no empirical support by demonstrating that the studies have been inconclusive precisely because key assumptions are yet to be isolated and clarified. Nevertheless, Pearl's admissions on the matter do place considerable weight

[220] The discussion of justifying modelling assumptions is somewhat continuous with the literature on justifying auxiliary hypotheses as well as the various issues that surround the problems and successes of the hypothetico-deductive method. See, for instance, Shipley (2000: 50). The discussion is also continuous with the literature on the nature of experiment in science.

on the need to discharge modelling assumptions before pronouncing causality. To reiterate the point, the issue of representing causal quantities requires, in addition to measures of identifiability, the justification of modelling assumptions within the context of the domain the models serve. Whilst Pearl (2000a) mentions the issue only in passing, elsewhere Pearl offers the following comment:

> Causal analysis with graphical models does not deal with defending modelling assumptions, in much the same way that differential calculus does not deal with defending the physical validity of a differential equation that a physicist chooses to use. In fact no analysis void of experimental data can possibly defend modelling assumptions. Instead causal analysis deals with the conclusions that logically follow from the combination of data and a given set of assumptions, just in case one is prepared to accept the latter. Thus, all causal inferences are necessarily *conditional*. These limitations are not unique to graphical models. In complex fields like the social sciences and epidemiology, there are only few (if any) real life situations where we can make enough compelling assumptions that would lead to the identification of causal effects (Freedman 2002: 15).

There are two initial lessons to be drawn. First, it is improper to provide a model with a causal interpretation in ignorance of any accompanying modelling assumptions and without acknowledging the procedure followed by an investigator to vet modelling assumptions. Second, the vetting of modelling assumptions will vary from model to model and from domain to domain[221]. Both lessons point toward the importance of recognising how models are used in science and are suggestive of a considerably broad methodological pluralism between some branches of science.

But on my reading, these are lessons Pearl (2000a) has apparently learnt well, a fact that travels some way towards explaining why Pearl (2000a) contains almost no

direct discussion of modelling assumptions. Indeed, Freedman's (2002) assertions concerning the dangers of inferring causality based solely on the causal modelling methods of Pearl's regimentation and formalism are puzzling when read next to Freedman's acknowledgment that Pearl admits all causal inferences are necessarily conditional. The composition of Pearl's theory (as I have interpreted it) re-enforces the point. The division of labour between the two parts assigns the objective account–not the regimentation–the task of detailing a basic set of conditions that set the standard which modelling assumptions should meet if they are to justifiably represent cause/effect relationships. For Pearl, it is just these standards that allow one to judge when experimental techniques discern causal relationships. Hence, the attempt to establish that the first part of Pearl's theory is open to problems of identification and model justification is not sufficient reason from which to conclude that Pearl's theory does not succeed in modelling causal relationships.

Pearl (2004) provides an indication of how to assess whether the causal relationships implied by a given model are in fact correct. Pearl considers that the issue of justifying modelling assumptions amounts to the problem of assessing whether the causal assumptions that support a given interpretation of a model actually hold in the real world. As is clear from the citations of this section, Pearl admits that such assumptions are primarily based on human judgement and cannot generally be tested unless those assumptions impose constraints on the model. Here Pearl finds that what is required is a formal method of assessing to what degree a given interpretation is *robust* to violations of those assumptions, since an interpretation that is robust to violations of model assumptions renders that interpretation more credible than one that is sensitive to the causal assumptions underlying a model (Pearl 2004: 1). The intuition behind this claim is that an estimate of a causal quantity is correct or, at least, highly plausible, when one can arrive at it via more than one means of calculation so long as each means of calculation employed is distinct from the

[221] i.e. from science to science if not also from study to study. For instance, the defence of modelling assumptions for causal models of economic phenomena will vary considerably from assumptions taken in epidemiology, psychology, and so forth.

others[222] (Pearl 2004: 2). For Pearl, this intuition finds expression via the notion of testability:

> It is only through violating [a model's] implied constraints that we can falsify a model, and it is only by escaping the threat of such violation that a model attains our confidence, and we can then state that the model and some of its implications (or *claims*) are *corroborated* by the data (Pearl 2004: 2).

So it would seem that Pearl's regimentation and formalism can take one a considerable way towards assessing whether or not a set of causal claims from a specific model are correct. According to Pearl it is possible in at least some instances both to make clear causal claims and to show those claims to be robust either by a process of reason or by demonstration. Freedman's prior criticism had been that one cannot pronounce causality without having first discharged the assumptions underlying a model, but that it is rarely the case that one has any compelling reasons for doing so. Whilst this may be true the forgoing discussion has shown that both Freedman and Pearl agree that where the requisite assumptions have been discharged one may proceed to draw causal inferences from a model.

One might wish to interject that these responses only serve to relocate problems of interpretation and justification from Pearl's regimentation to Pearl's objective account. Indeed, I have already expressed my view that Pearl's objective account is the correct place to deal with such problems. It is, therefore, appropriate to question whether it follows, even granted a well-identified model with a robust set of causal claims, that the model actually succeeds in representing causal relationships present in the system under study[223]. For it would appear that nothing said so far rules out the possibility that a well-identified and apparently well-justified model based on Pearl's regimentation may nevertheless fail to capture (all relevant) actual causal relationships present in the system under investigation, even granted such actual

[222] Where two methods are considered distinct so long as neither shares the same assumptions

causal relationships fit the definition of objective constraints specified by Pearl's objective account.

The mere possibility is enough reason on which to doubt the validity of the inference to causality[224]. Put another way, what one aims to do when interpreting and justifying causal models is to answer the following questions:

(Q1) *Interpretation* 'What does the model say about the way the world is?'

(Q2) *Justification* 'Is the world the way the model says it is?'

According to the present objection, Pearl can answer Q1 readily but either fails to answer Q2 at all or provides answers that are (often) false[225]. Although I think this complaint may gain traction it is difficult to assess constructively in the absence of close argument about actual models of real data[226]. However, the general charge is one that seeks to question the credentials a model can have to represent causality. Some discussion of how a model might attain such 'credentials' is in order.

It has been accepted for some time that causal models do not really succeed in definitively answering questions such as Q2. For instance, Koopmans admits:

> [...] the research worker who constructs a model does not really believe that reality is exactly described by a 'true' structure contained in the model. Linearity, discrete time lags, are obviously only approximations. At best, the model builder hopes to build a model that contains a structure that

---

[223] That is, whether the model structure correctly represents the modelled system. Cartwright (2003) expresses reservations.

[224] Of course, the strength of this assertion depends on the quality of counterexamples set out against it. Cartwright (2003: 258-265) provides a sketch of such counterexamples.

[225] In support of an equivalent claim Cartwright (2003) asserts that in many of the domains Pearl's theory is intended to be applied, we should expect Pearl's models to answer Q2 only rarely since the causal connections common in those domains are of a sort not readily detectable by Pearl's theory.

[226] For example, recall Bouman's point that models can be data probes and consequently may not be cleanly separable from the system they model.

> approximates reality to a degree sufficient for the practical purposes of the investigation. [The research worker needs] to choose the simplest possible set (from two or more possible sets of structures) – in some sense – that contains a structure sufficiently approximate – in some sense – to [...] reality (Koopmans 1985: 121).

From comments I cite above it is clear that Pearl accepts Koopmans point also. We can grant then, that the complete justification of a causal model remains an ideal[227]. But this immediately begs the question of how one decides whether and in what sense a causal model is 'sufficiently approximate' to reality. Presumably, a good answer to this question must provide some assurance that the model has captured all relevant causal relationships present in the underlying system. The credentials of the model to make causal claims will depend on the quality of such assurances. But, as Winsberg (2003) points out, judging a model's credentials is not a simple process. A model's credentials do not straightforwardly follow from the model's fidelity to theory or to data or by the application of some *a priori* standard (Winsberg 2003: 121)[228]. Instead, credentials are earned in the field against numerous competing commitments:

> Whenever [modelling] techniques and assumptions are employed successfully, that is, whenever they produce results that fit well into the web of our previously accepted data, our observations, the results of our paper and pencil analyses, and our physical intuitions, whenever they make successful predictions or produce engineering accomplishments, their credibility as reliable techniques or reasonable assumptions grows (Winsberg 2003: 122).

---

[227] But, see Giere (forthcomingb) for an alternative approach to assessing modelling assumptions and claims.

[228] See also Pearl (2001: 3). I note in passing that many comments throughout Pearl (2000a) suggest commitment to the view that models may be tested and falsified against observation, a view parallel to the one held by some philosophers that observation and experiment provide a clear view of natural facts and regularities against which theories may be tested. Commitment to such a position seems at odds with the aims of Pearl (2004) regarding robustness since the latter appears to rule out model falsifiability. Indeed, Pearl's views on the testability of modelling assumptions by experiment, why experimental methods reveal causal mechanisms, and robustness, when taken together as a set appear inconsistent. But, see discussion of related issues by Giere (forthcominga: 17). For further discussion

The credibility of a model is not, therefore, an all or nothing matter and, even in the absence of outstanding accomplishments, modelling techniques and assumptions may still produce informative results[229]. Therefore, Pearl's theory does not stand or fall depending on whether or not its techniques capture the actual causal relationships of a system under investigation, and so Pearl is spared the need to make a new case for the efficacy of his theory. In fact, the key elements of how the theory responds to questions of model justification are already on the table and may be summarised in the following way.

Recall that, for Pearl, studies by the behavioural, social and bio-medical sciences describe real systems in terms of random variables and presuppose that some of these random variables have a causal influence on others (Halpern and Pearl 2001a). The investigators who carry out such studies aim to locate and to describe the nature of these influences. The standard approach to locating which variables have a causal influence is to construct an appropriate experiment. Pearl asserts that the scientific experiment has two important components. The first involves intervention, which consists in circumscribing the real system of interest and the second involves randomisation, which renders the values of the variables of the circumscribed system either due to chance or to persistent influences still present in the circumscribed system. Logic is then applied to the results to determine the reason for the persistent influences[230]. Recall from the discussion of section 1.2 that at first glance the explanation of the persistent influences amongst experimental variables must be one of three types: either one variable is the cause of the other, or vice versa, or there are common causes influencing both (Shipley 2000: 8). The further one strays from the

---

of the relationship between autonomy and falsification consistent with Pearl (2004) see Boumans (2003).

[229] See discussion of related matters by Sklar (2003).

[230] Further intervention and observation may also follow.

spirit of such experimental techniques the less one is able to infer causality with any precision[231].

Further recall that, for Pearl, it is not experimental design in-and-of-itself that licenses causal inference. Instead the applicability of experimental techniques to problems involving causality follow from the fact that such techniques offer a way of locating what Pearl describes as autonomous mechanisms. In contrast to the experimenter, the (non-experimental) modeller seeks to construct an assembly of functional mechanisms that can explain how the data taken from the real system were produced without recourse to the actual realisation of a specific experimental set-up. The modeller then makes causal claims on behalf of the mathematical details and assumptions of the model rather than on behalf of the experimental design. The ground common to both experimental and non-experimental techniques, according to Pearl, is the attempt to locate and to describe causal influences present in real systems.

According to Pearl's objective account a relationship is causal if it involves objective constraints on physical processes. As I have discussed this is a fairly loose condition. For instance, it is not as detailed as Dowe's account of causal processes as world-lines of objects possessing physically conserved quantities (Dowe 2000: 90). What it does say is that causal relationships are mind-independent law-like constraints on interactions between objects or events.

I have drawn a connection between the autonomous mechanisms of Pearl's theory and the account of mechanism offered by Glennan (1996, 2000a, 2000b, 2002). As I have discussed, for Glennan a mechanism is a kind of complex system, which produces a number of behaviours. A mechanism for a specific behaviour produces that behaviour by the interaction of a number of parts characterised by direct,

[231] Of course, this is a simplification of experimental methods. For discussion of experiment in science see, for example, Hacking (1991), Franklin (1989) and the collection of essays edited by Radder (2003).

invariant and change-relating generalisations (Glennan 2002: S345)[232]. On Glennan's account, a mechanism may be decomposed into a number of parts. Parts, for Glennan must be objects rather than events. To be an object is to continue to possess properties in the absence of interventions. Parts are generally spatially localized. How a system is decomposed, and what parts result, depends upon which behaviour is being considered. An interaction is an occasion on which a change in a property of one part brings about a change in a property of another part. Direct, invariant, change-relating generalisations are equated with some number of counterfactual claims. A generalisation is 'change-relating' when it describes a relationship between two or more parts such that an intervention that changes one part will bring about a change in another part. A change-relating generalisation is 'invariant' when it remains true under some class of interventions performed on the background conditions of the mechanism described by the generalisation. And an invariant, change-relating generalisation is 'direct' when the generalisation pertains to the exclusive interactions between two parts, a condition which, according to Glennan (2000b), rules out generalisations that would truly describe relationships between two distinct parts in which one part indirectly caused a change in the properties of a second part by changing properties of one or more intervening parts (Glennan 2000b: 9-10). Finally, on Glennan's analysis of mechanism, a complex system is typically hierarchical in the sense that objects that are parts of a mechanism may themselves be complex mechanisms that can be decomposed into further parts (Glennan 2000b: 10).

Note that the key conditions Glennan specifies in his analysis of mechanism include the following notions: mechanisms are decomposable into parts; the interaction between parts is direct; an intervention on one part produces change in another, *bona fide* mechanisms display a sort of equilibrium other relationships are missing; mechanical relationships remain stable under a range of interventions; mechanisms can be described at different levels of abstraction. As I claimed in chapter 1, the reason that Glennan makes for an excellent philosophical touchstone for interpreting

[232] The term 'direct, invariant, change-relating generalisation' replaces Glennan's earlier use of the term 'direct causal laws'.

Pearl's theory of causality is that Pearl accepts (a version) of each of these notions. Glennan's notion of decomposition, directness, abstraction, and stability are reflected in the conjunction of Pearl's determinism, causal Markov, modularity, minimality and stability conditions. Moreover, Glennan's idea that the relationships between the parts of a mechanism are change relating is a counterpart of Pearl's notion of manipulation. Similarly, for Glennan, mechanisms are the entities we appeal to when explaining the behaviour of real world systems; a causal relation between two events exists by virtue of the mechanism that connects them (Glennan 1996: 64). On the latter point Glennan is also in agreement with Salmon that it is causes that underwrite explanations rather than explanations underwriting causes. But, since explanations are for Glennan an appeal to mechanisms conceived as descriptions of the behaviours of complex systems, Glennan's account is an ontic conception of mechanism. This means that each of the notions that together form Glennan's analysis of mechanism is to be understood as a description of certain features of a mind-independent reality.

This is not so for Pearl (2000a). On my interpretation of Pearl's theory, the notions stand as regimented causal concepts put to work in a formal calculus, each a part of an apparatus designed for detecting causal relationships rather than the result of an analysis of causal relationships. Hence, the conditions are, according to Pearl's theory, epistemic rather than ontic. What one aims to detect or delimit via the employment of these notions are, according to the objective account, objective constraints on physical processes.

3.5 Concluding Remarks

According to my interpretation Pearl (2000a) provides a theory of causality in two parts. I have described one part of the theory as an objective account of causation and the other part as the regimentation of a number of causal concepts important to discovery and explanation in the special sciences. The key features of Pearl's theory of causality are a formal language purpose built for expressing causal claims, a statement about the relationship between causal claims and objective constraints on real systems, and a logic for guiding inference with causal claims. I have interpreted Pearl's theory in two parts principally on the basis that Pearl's comments about causality are of two kinds: those that reflect the use of causal terms and concepts in the sciences on the one hand and those that address what (Pearl thinks) causal relationships *are* on the other. That is, the formal component of Pearl's theory encodes causal concepts and, reflected in statements about how causal claims relate to objective constraints, is an account of what makes causal relationships causal. The fact that Pearl spends the greater amount of time on the detail of the regimentation and less time elaborating the objective account is evidence that Pearl's primary interest in causality lay in articulating a theory of causal modelling rather than in analysing causation. Pearl's interest in causal modelling over analysis rendered the natural option of interpreting Pearl's theory into the language of mainstream philosophy of causation unattractive. Instead, I identified as the appropriate interpretive medium research in the philosophy of science that aims to understand the nature and use of scientific models.

In chapter 2 I presented what I consider to be the standard features of Pearl's theory. I then focussed on the theory's use of counterfactuals, the theory's demarcation between statistical and causal information and concepts, and the theory's method of justifying causal claims. I think these components hold the greatest importance for understanding Pearl's theory. My examination of each component revealed: (i) the theory's employment of counterfactuals is devised to encode the sorts of conditions scientists decide are important when reasoning about modality in experimental

settings and simulations; (ii) the theory exemplifies the view that causal claims cannot be expressed in the language of probabilities and that it is causal relationships which drive the important associations described by statistics; (iii) the truth of a model's claims about causal relationships are judged not by demonstration or by showing the model's assumptions to hold in the system under study, but via a vindication of the model's claims in spite of the model's assumptions.

According to my interpretation of Pearl's theory the thrust of the regimentation is epistemic and the thrust of the objective account is ontic. The regimentation pertains to the discovery of causal relationships and the objective account purports to provide an objective analysis of causation. On examination the objective account was found to provide only a general sketch of causation, one that says so little it almost defies categorising. The objective account allows that there are causal relationships beyond those detectable via the employment of the regimentation. Given the regimentation is not designed to uncover every instance of what passes for a causal relationship according to the objective account, the regimentation is one approach to causal modelling among many candidates. This stands to reason since there are, after all, many different ways of representing objective constraints in the sciences. I have suggested that what the objective account says is reminiscent of a nomic account of causal processes but that the level of detail precludes identifying where the account stands, for instance, in relation to the Humean supervenience thesis. Given the efficacy of the two-part interpretation, it is interesting to speculate about why so little detail is afforded to the objective account, doubly so when one considers the amount of time philosophers spend analysing causation. No doubt the answer is reflected in the aims and commitments of the causal discovery programs from which Pearl's theory springs. On reflection perhaps the most important part of Pearl's theory is its approach to the study of causality. According to the theory, the description of a causal connection is bound to the methods employed in its discovery. This means that many disagreements about the nature of causal relationships may be misunderstandings that can be avoided by adopting a common language for causal talk. These statements have a positivist ring to them and deserve closer examination. Indeed, I have argued

for the soundness of the interpretation while reserving critical evaluation of the result for a separate project. But, even in the absence of a critical evaluation it is clear that Pearl's theory, *contra* Freedman, says nothing about causality which is at odds with elementary scientific principles in an obvious way. Experimental and non-experimental methods are alike in that both presuppose a model of causal discovery often prior to, if not always in conjunction with investigation of what the world is like. Finally, I have said that the case for categorising Pearl's theory as a manipulability or as a counterfactual theory of causality is not clear cut. Although it is clear that Pearl's theory makes mention of notions related to both, this does not seem enough to commit the theory to either. In light of my interpretation I suggest that the arguments made for such a classification of Pearl's theory face a dilemma: they either (i) confuse the nature of the causal relationships identified via application of the theory with the tools employed in the application of the theory, or (ii) can produce evidence from Pearl (2000a) for a more elaborate objective account than the one I have described above. Of course, all these claims are controversial and cannot be the last word. Pearl's theory continues to be developed and debate proceeds about what role key notions such as invariance have to play[233]. The critical evaluation should prove exciting.

[233] For example, see Woodward (2000, 2002a, 2002b), Menzies (2002), Cartwright (2000b) and Boumans (2003).

References:

Adams, E.W., (1975) *The Logic of Conditionals*, Dordrecht: Reidel.

Anscombe, G.E.M., (1993) "Causality and Determination," in E. Sosa and M. Tooley (eds.), (1993) *Causation*. Oxford Readings in Philosophy (ser.), New York: Oxford University Press (pp 88-104).

Armstrong D., (1997) "Singular Causation and Laws of Nature" in *The Cosmos of Science*, J. Earman and J. Norton (eds.), Pittsburgh: Pittsburgh University Press, (498-511).

Armstrong, D., (1999a) *A World of States of Affairs*. Cambridge: Cambridge University Press.

Armstrong, D., (1999b) "The Open Door: Counterfactual Versus Singularist Theories of Causation" in H. Sankey (ed.), *Causation and Laws of Nature*. Dordrecht: Kluwer (175-185).

Aronson, J., (1971) "The Legacy of Hume's Analysis of Causation" *Studies in the History and Philosophy of Science* (2: 135-156).

Austin, J.L., (1961) *Philosophical Papers*, (ed.) J.O. Urmson and G.J. Warnock, Oxford: Oxford University Press.

Barker, S.J., (1995) "Towards a Pragmatic Theory of 'If'" *Philosophical Studies* (79: 185-211).

Barker, S.J., (1991) "*Even, Still*, and Counterfactuals" *Linguistics and Philosophy* (14: 1-38).

Beauchamp, T. and N. Rosenberg, (1981) *Hume and the Problem of Causation*. New York: Oxford University Press.

Berkovitz, J., (2002) "On Causal Inference in Determinism and Indeterminism", in Atmanspacher, H. and R. Bishop (eds.), (2002) *Between Chance and Choice: Interdisciplinary Perspectives on Determinism*, Thorverton; Imprint Academic, (pp 237-278).

Bigelow, J. and R. Pargetter, (1990) *Science and Necessity*. Cambridge: Cambridge University Press.

Blalock, H. M. Jr., (ed.), (1985) *Causal Models in the Social Sciences*. New York: Aldine Publishing Company (2nd edition.).

Bollen, K.A., (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Boumans, M., (2001) "Measure for Measure: How Economists Model the World into Numbers" *Social Research* (68: 427-453).

Boumans, M., (2002) "Calibration of Models in Experiments." In Magnani, L. and N.J. Nersessian (eds.), *Model-Based Reasoning: Technology, Science, Values*, New York: Kluwer Academic/Plenum Publishers, (75-93).

Boumans, M., (2003) "How to design Galilean Fall Experiments in Economics." *Philosophy of Science* (70: 2: 308-329).

Braithwaite, R.B., (1953) *Scientific Explanation*, Cambridge: Cambridge University Press.

Cartwright, N., (1979) "Causal Laws and Effective Strategies" *Nous* (13: 419-437).

Cartwright, N., (1995) "False Idealisation: A Philosophical Threat to Scientific Method," *Philosophical Studies* (77: 339-352).

Cartwright, N., (2000) "Causal Diversity and the Markov Condition," *Synthese* (121: 3-27).

Cartwright, N., (2001) "Modularity: It can–and Generally Does–Fail", in Galavotti, M., P. Suppes and D. Constantini (eds.) *Stochastic Causality*, Stanford: CSLI Publications.

Cartwright, N., (2002a) "Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward" *British Journal for the Philosophy of Science* (53: 411-453).

Cartwright, N., (2002b) "Two Theorems on Invariance and Causality" *Philosophy of Science* (70: 203-224).

Cartwright, N., (2003) "What is Wrong with Bayes Nets?" In H.E. Kyburg Jr. and M. Thalos (Eds.) *Probability Is the Very Guide to Life*, Chicago: Open Court, (pp 253-275).

Cartwright, N., (forthcoming) "Two Concepts of Invariance, Two Concepts of Intervention and Two Concepts of What It Is to Be Causally Correct" *Philosophy of Science.*

Charniak, E., (1991) "Bayesian Networks without Tears" in *AI Magazine* (Winter).

Cole, S.R. and M.A. Hernan, (2002) "Problems with Estimating Causal Effects." *International Journal of Epidemiology* (31:1: 163-165)

Cook, T. and D. Campbell, (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Miflin Company.

Copeland, B.J., (1983) "Pure Semantics and Applied Semantics" *Topoi* (2: 197-204).

Cosmides, L. and J. Tooby, (1996) "Are Humans Good Intuitive Statisticians After all? Rethinking Some Conclusions from the Literature on Judgement Under Uncertainty" *Cognition* (58: 1-73).

Darden, L., (2002) "Strategies for Discovering Mechanisms" *Philosophy of Science* (69: S354-S365).

Dawid, A.P., (1979) "Conditional Independence in Statistical Theory" *Journal of the Royal Statistical Society Ser. B* (41: 1-31).

Dawid, A. P., (2000) "Causal Inference without Counterfactuals", *Journal of the American Statistical Association* (95: 407-448).

De Finetti, B., (1990) *Theory of Probability (Vol I, II)*. Chichester: Wiley Classics Library, John Wiley and Sons.

Dodge, Y., (ed.), (2003) *The Oxford Dictionary of Statistical Terms*, New York: Oxford University Press.

Dowe, P., (2000) *Physical Causation*, Cambridge: Cambridge University Press.

Duhem, P., (1954) *The Aim and Structure of Physical Theory*. P. Weiner (trans.) Princeton: Princeton University Press.

Duncan, O.D., (1975) *Introduction to Structural Equation Models*. New York: Academic Press.

Edgington, D., (1995) "On Conditionals" *Mind* (104: 235-329).

Edwards, D., (2000) *Introduction to Graphical Modelling*, (2nd edition) Springer.

Eells, E., (1991) *Probabilistic Causality*. Cambridge: Cambridge University Press.

Ehring, D., (1998) *Causation and Persistence*. Oxford: Oxford University Press.

Fodor, J.A., (1974) "Special Sciences, or The Disunity of Sciences as a Working Hypothesis" *Synthese* (28: 97-115).

Franklin, A., (1989) *The Neglect of Experiment*. Cambridge: Cambridge University Press.

Freedman, D.A., (1997) "From Association to Causation via Regression – with disucssion," in V. McKim and S. Turner, (eds.), *Causality in Crisis?* Indiana: University of Notre Dame Press, (pp 113-182).

Freedman, D.A., (2002) "On Specifying Graphical Models for Causality and the Identification Problem" Technical Report #601, (http://stat-www.berkeley.edu/~census/601.pdf).

Freedman, D.A., (2003) "From Association to Causation: Some Remarks on the History of Statistics," in *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century*. J. Panaretos (ed.) (45-71).

Freidman, M., (1974) "Explanation and Scientific Understanding" *Journal of Philosophy* (71: 5-19).

Gardiner, P., (1959) *The Nature of Historical Explanation.* Oxford: Oxford University Press.

Gasking, D., (1955) "Causation and Recipes" *Mind* (64: 479-487).

Geiger, D., T. Verma, and J. Pearl, (1990) "Identifying Independence in Bayesian Networks" *Networks* (20: 507-534).

Gibbard, A., (1981) "Two Recent Theories of Conditionals," in W. Harper, Stalnaker and Pearce (eds.), *Ifs*. Dordrecht: Reidel.

Giere, R.N., (1999) "Using Models to Represent Reality," in Magnani, L., N. Nersessian and P. Thagard, (eds.), *Model-Based Reasoning in Scientific Discovery.* New York; Kluwer Academic/Plenum Publishers, (41-57).

Giere, R.N., (forthcominga) "Perspectival Pluralism" *Minnesota Studies in the Philosophy of Science.*

Giere, R.N., (forthcomingb) "How models are used to represent reality" *Philosophy of Science.*

Gillies, D., (2001) "Critical Notice: Judea Pearl Causality: Models, Reasoning, and Inference" *British Journal for the Philosophy of Science* (52: 613-622).

Glennan, S.S., (1996) "Mechanisms and the Nature of Causation" *Erkenntnis* (44: 49-71).

Glennan, S.S., (2000a) 'A Model of Models' manuscript.

Glennan, S.S., (2000b) 'Rethinking Mechanistic Explanation.' Manuscript available at URL = <http://hypatia.ss.uci.edu/lps/psa2/program.html.> (version as at 21/12/00).

Glennan, S.S., (2002) "Rethinking Mechanistic Explanation" *Philosophy of Science* (69: S342-S353).

Glock, H-J., (1996) *A Wittgenstein Dictionary*. Oxford: Blackwell Publishers.

Glymour, C., (2003) "Instrumental Probability," in H.E. Kyburg, Jr., and M. Thalos (eds.) *Probability is the Very Guide of Life*. Chicago: Open Court, (pp 235-252).

Greenland, S., and B. Brumback (2002) "An Overview of Relations Among Causal Modelling Methods" *International Journal of Epidemiology* (31: 5: 1030-1037).

Grice, H.P., (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haack, S., (1978) *Philosophy of Logics*. Cambridge: Cambridge University Press.

Haavelmo, T., (1943) "The Statistical Implications of a System of Simultaneous Equations" *Econometrica* (11: 1-12).

Hacking, I., (1965) *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, I., (1991) *Representing and Intervening*. Cambridge: Cambridge University Press.

Hajek, A., (2003) "Conditional Probability Is the Very Guide of Life" in H.E. Kyburg, Jr., and M. Thalos, (eds.), (2003) *Probability Is the Very Guide of Life*. Chicago: Open Court (pp 183-203).

Halpern, J. and J. Pearl, (2001a) "Causes and Explanations: A Structural-Model Approach – Part I: Causes," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufman (194-202).

Halpern, J. and J. Pearl, (2001b) "Causes and Explanations: A Structural-Model Approach – Part II: Explanations," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufman.

Hankinson Nelson, L., (2003) "The Descent of Evolutionary Explanations: Darwinian Vestiges in the Social Sciences," in *The Blackwell Guide to the Philosophy of the Social Sciences*, S.P. Turner and P.A. Roth, (eds.), Cornwall: Blackwell Publishing (258-289).

Hart, H. and T. Honore, (1985) *Causation in the Law*. Oxford: Clarendon Press.

Hausman D. and J. Woodward, (1999) "Independence, Invariance, and the causal Markov condition", *British Journal for the Philosophy of Science* (50: 4: 521-583).

Heckman, J., (2001) "Causal Parameters and Policy Analysis in Econometrics: A Twentieth Century Retrospective" *The Quarterly Journal of Econometrics* (CVX: 45-97).

Heckerman, D., C. Meek, and G. Cooper, (1999) "A Bayesian Approach to Causal Discovery." In C. Glymour and G. Cooper, (eds.), *Computation, Causation, and Discovery*, Cambridge, MA: MIT Press, (pp. 143-167).

Hedstrom, P., and R. Swedberg, (eds.) (1998) *Social Mechanisms*. Cambridge: Cambridge University Press.

Hempel, C., (1942) "The Function of General Laws in History." *Journal of Philosophy* (39: 35-48).

Hempel, C., (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Hempel, C., and P. Oppenheim, (1948) "Studies in the Logic of Explanation." *Philosophy of Science* (15: 135-175).

Herfel, W., W. Krajewski, I. Niiniluoto, and R. Wojcicki, (eds.) (1995) *Theories and Models in Scientific Processes*. Poznan Studies in the Philosophy of the Sciences and the Humanities, 44 (Ser.), Amsterdam: Rodopoi.

Hertz, H., (1894) *The Principles of Mechanics*. D.E Jones and J.T. Walley (trans.) London: Macmillan.

Hitchcock, C., (2001) "Book Review - Causality: Models, Reasoning and Inference by Judea Pearl" *The Philosophical Review* (110: 4: 639-641).

Hitchcock, C. and J. Woodward, (2003a) "Explanatory Generalizations, Part I: A Counterfactual Account" *Nous* (37: 1: 1-24).

Hitchcock, C. and J. Woodward, (2003b) "Explanatory Generalizations, Part II: Plumbing Explanatory Depth" *Nous* (37: 2: 181-199).

Hodges, W., (2001) "Model Theory." *The Stanford Encyclopaedia of Philosophy (Winter Edition)*, Edward, N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2001/entries/model-theory/>.

Hoover, K.D., (2003) "Book Review: J. Pearl, Causality" *The Economic Journal* (113: 488: F411-F413).

Hopkins, M. and J. Pearl, (2003) "Clarifying the Usage of Structural Models for Commonsense Causal Reasoning," in *Proceedings of the AAAI Spring Symposium on the Logical Formalization of Commonsense Reasoning*, Menlo Park, CA: AAAI Press (83-89).

Horwich, P., (1982) *Probability and Evidence*. Cambridge: Cambridge University Press.

Horwich, P., (1987) *Asymmetries in Time*. Cambridge: MIT Press.

Horwich, P., (1990) *Truth*. Oxford: Blackwell.

Howson, C. and P. Urbach, (1993) *Scientific Reasoning: The Bayesian Approach*. (Second edition), Chicago: Open Court.

Humphreys, P., (1997) "A Critical Appraisal of Causal Discovery Algorithms." In V. McKim and S. Turner (eds.), *Causality in Crisis?* South Bend, IN: University of Notre Dame Press (249-263).

Humphreys, P., (1989) *The Chances of Explanation*. Princeton University Press: Princeton.

Humphreys, P. and D. Freedman, (1996) "The Grand Leap" *British Journal for the Philosophy of Science* (47: 113-123).

Humphreys, P. and D. Freedman, (1999) "Are There Algorithms that Discover Causal Structure?" *Synthese* (121: 29-54).

Jackson, F., (1979) "On Assertion and Indicative Conditionals", Philosophical Review (88: 565-589).

Jackson, F., (ed.), (1991) *Conditionals*, Oxford: Oxford University Press.

Jackson, F., (1994) "Armchair Metaphysics" in Michael, M. and J. O'Leary-Hawthorne (eds.) *Philosophy of Mind*, Dordrecht: Kluwer, (23-42).

Jackson, F., (1998) *From Ethics to Metaphysics*. Oxford: Oxford University Press.

Jeffrey, R., (1964) "'If'" *Journal of Philosophy* (61: 702-703).

Jordan, M.I., (1998) *Learning in Graphical Models*, ser. D., vol. 89, (Behavioural and Social Sciences). Dordrecht: Kluwer.

Kennett, R.J., K. Korb, and A.E. Nicholson, (2001) "Seabreeze Prediction Using Bayesian Networks: A Case Study," in D. Cheung, G. Williams, and Q. Li, (eds.), *Advances in Knowledge Discovery and Data Mining: 5th Pacific-Asia Conference* (PAKDD), Hong Kong, China. Proceedings, Lecture Notes in Artificial Intelligence vol. 2035, (148-153).

Kitcher, P., (1976) "Explanation, Conjunction and Unification" *Journal of Philosophy* (73: 207-212).

Kitcher, P., (1981) "Explanatory Unification" *Philosophy of Science* (48: 251-281).

Kitcher, P., (1985) "Two Approaches to Explanation" *Journal of Philosophy* (82: 632-639).

Kitcher, P., (1989) "Explanatory Unification and Causal Structure" *Minnesota Studies in the Philosophy of Science*, 13, Minneapolis: University of Minnesota Press, (pp 410-505).

Kline R.B., (1998) *Principles and Practice of Structural Equation Modelling*. New York: Guilford.

Koopmans, T.C., (1950) "When is an equation system complete for statistical purposes?" In T.C. Koopmans, (ed.) *Statistical Inference in Dynamic Economic Models*. New York: Wiley.

Koopmans, T. C. (1985) "Identification Problems in Economic Model Construction", in Blalock, H. M. Jr., (ed.) *Causal Models in the Social Sciences*. New York: Aldine Publishing Company (Second edition), (pp 103-123).

Korb, K.B. and C.S. Wallace, (1997) "In Search of the Philosopher's Stone: Remarks on Humphreys and Freedman's Critique of Causal Discovery" *British Journal for the Philosophy of Science* (48: 543-553).

Kvart, I., (1986) *A Theory of Counterfactuals*. Indianapolis: Hackett.

Kyburg, H., (1974) *Logical Foundations of Statistical Inference*. Boston: Reidel.

Kyburg, H.E., Jr. and Smokler, H.E., (eds.) (1980) *Studies in Subjective Probability*. (Second edition), New York: John Wiley and Sons.

Lakatos, I., (1970) "Falsification and the Methodology of Scientific Research Programmes," in I. Lakatos and A. Musgrave, (eds.) *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press (91-195).

Laplace, P.S., (1814) *Essai philosophique sur les probabilities*. Paris: Courcier. Reprinted (1912) F.W. Truscott and F.L Emory (Trans.), Wiley: New York.

Lemmer, J.F., (1993) "Causal Modelling," in D. Heckerman and A. Mamdani, (eds.) *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*. San Mateo CA: Morgan Kaufmann (143-151).

LeRoy, S.F., (2002) "Review of Causality: Models, Reasoning and Inference, by Judea Pearl" *Journal of Economic Methodology* (9: 1: 101-103).

Levi, I., (1977) "Direct Inference" *Journal of Philosophy* (74: 5-29).

Levi, I., (2003) "Objective Modality and Direct Inference" in H.E. Kyburg, Jr., and M. Thalos, (eds.) (2003) *Probability Is the Very Guide of Life*. Chicago: Open Court (pp 61-87).

Lewis, D., (1973a) "Counterfactuals and Comparative Possibility" In W.L. Harper, R. Stalnaker, G. Pearce (eds.) *Ifs*. Dordrecht: Reidel, (pp 57-85).

Lewis, D., (1973b) *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1979) "Counterfactual Dependence and Time's Arrow" *Nous* (13: 418-446).

Lewis, D., ([1980] 1986) "A Subjectivist's Guide to Objective Chance," in D. Lewis, *Philosophical Papers Volume II*. New York: Oxford University Press.

Lewis, D., (1986) *Philosophical Papers Volume II*. New York: Oxford University Press.

Lycan, W.G., (2001) *Real Conditionals*. Clarendon Press: Oxford.

Mackie, J.L., (1974) *The Cement of the Universe* Oxford: Oxford University Press.

Machamer, P., (2002) "Activities and Causation: the Metaphysics and Epistemology of Mechanisms," paper presented at the Eighteenth Biennial Meeting of the PSA, Milwaukee, WI.

Machamer, P., L. Darden, and C. Craver, (2000) "Thinking about Mechanisms" *Philosophy of Science* (67: 1-25).

Magnani, L., and N.J. Nersessian, (eds.) (2002) *Model-Based Reasoning: Technology, Science, Values*. New York: Kluwer Academic/Plenum Publishers.

Magnani, L., N. Nersessian and P. Thagard, (eds.) (1999) *Model-Based Reasoning in Scientific Discovery*. New York; Kluwer Academic/Plenum Publishers.

Maldonado, G. and S. Greenland, (2002) "Estimating Causal Effects (with discussion)" *International Journal of Epidemiology* (31: 2: 422-438).

Martel, I., (forthcoming) "The Principle of Common Cause, the Causal Markov Condition, and Quantum Mechanics: Comments on Cartwright." In *Nancy Cartwright's Philosophy of Science*, S. Hartmann and L. Bovens, (eds.).

McKim, V., and S. Turner, (eds.) (1997) *Causality in Crisis?* South Bend, IN: University of Notre Dame Press.

Meheus, J., (ed.) (2002) *Inconsistency in Science*. Dordrecht: Kluewer Academic Publishers.

Menzies, P., (1999) "Intrinsic versus Extrinsic Conceptions of Causation" in H. Sankey (ed.) *Causation and Laws of Nature*, Dordrecht: Kluwer (313-330).

Menzies, P., (2002) "Causal Models, Token-Causation and Processes," manuscript available at http://www.philsci-archive.pitt.edu/archive/00001039/00/PSA2002_long.pdf (6/11/03).

Menzies, P. and H. Price, (1993) "Causation as a Secondary Quality" *The British Journal for the Philosophy of Science* (44: 197-203).

Morgan, S.L., (2004) "Book Review: J. Pearl, Causality" *Sociological Methods and Research* (32: 3: 411-422).

Morrison, M., and M. Morgan, (eds.) (1999) *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Murphy, K., (2001) 'A Brief Introduction to Graphical Models,' online manuscript available at http://www.cs.berkeley.edu/~murphyk//Bayes/bayes.html, (Last accessed 26/9/02).

Nagel, E., (1961) *The Structure of Science: Problems in the Logic of Scientific Explanation*, New York: Harcourt, Brace and World.

Neuberg, L.G., (2003) "Book Review: J. Pearl, Causality" *Economic Theory* (19: 675-685).

Norton, S. and F. Suppe, (2001) "Why Atmospheric Modeling is Good Science", in C. Miller and P. Edwards, (eds.) *Changing the Atmosphere: Expert Knowledge and Environmental Governance*. Cambridge, MA: MIT Press.

Pearl, J., (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufman.

Pearl, J., (1995) "Causal Diagrams for Empirical Research" *Biometrika* (82: 669-710).

Pearl, J., (2000a) *Causality: Models, Reasoning, and Inference*. Cambridge University Press: New York.

Pearl, J., (2000b) "The Logic of Counterfactuals in Causal Inference (Discussion of 'Causal Inference Without Counterfactuals' by A.P. Dawid)," *Journal of the American Statistical Association* (95: 450: 428-435).

Pearl, J., (2001a) "Bayesianism and Causality, or, Why I am Only Half-Bayesian," in D. Corfield and J. Williamson, (eds.) *Foundations of Bayesianism*, Applied Logic Series Volume 24, Netherlands: Kluwer Academic Publishers (19-36).

Pearl, J., (2001b) 'D-separation without tears' URL = www.bayes.cs.ucla.edu/Book-2K/d-sep.html.

Pearl, J., (2002a) "Comments on Seeing and Doing" *International Statistical Review* (70: 2: 207-209).

Pearl, J., (2002b) "Comments on Nozer Singpurwalla's 'On Causality and Causal Mechanisms'," *International Statistical Review* (70: 2: 210-212).

Pearl, J., (2002c) "Causal Inference in the Health Sciences: A Conceptual Introduction" *Health Services and Outcomes Research Methodology* (2: 189-220).

Pearl, J., (2003a) "Statistics and Causal Inference: A Review" Test Journal (12: 2: 281-345).

Pearl, J., (2003b) "Comments on Woodward's 'Invariance, Modularity, and All That: Cartwright on Causation'," paper presented at *Nancy Cartwright's Philosophy of Science: An International Workshop*. University of Konstanz.

Pearl, J., (2004) "Robustness of Causal Claims," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Arlington, VA: AUAI Press (446-453).

Pearl, J. and A. Paz, (1987) "Graphoids: A Graph-Based Logic for Reasoning About Relevance Relations," in B. Du Boulay *et al.*, (eds.) *Advances in Artificial Intelligence, Vol. II*. Amsterdam: North-Holland (537-563).

Plantinga, A., (1974) *The Nature of Necessity*, Oxford: Oxford University Press.

Popper, K., (1959) *The Logic of Scientific Discovery*, London: Hutchinson.

Pratt, J. and R. Schlaifer, (1988) "On the Nature and Discovery of Structure" *Journal of the American Statistical Association* (79: 9-21).

Price, H., (1996) *Time's Arrow and Archimedes' Point: New directions for the physics of time*. New York: Oxford University Press.

Priest, G., (2001) *An Introduction to Non-Classical Logic*. Cambridge: Cambridge University Press.

Psillos, S., (2002) *Causation and Explanation*. Central Problems in Philosophy, John Shand, (Ser. ed.) Acumen: Bucks.

Putnam, H., (1973) "Reductionism and the Nature of Psychology" *Cognition* (2: 131-146).

Quine, W.V.O., (1980) "Two Dogmas of Empiricism," in *From a Logical Point of View*, (Second edition). Cambridge MA: Cambridge University Press (20-46).

Quine, W.V.O., (1995) *From Stimulus to Science*. Cambridge, MA: Harvard University Press.

Radder, H., (ed.) (2003) *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press.

Ramachandran, M., (1997) "A Counterfactual Analysis of Causation" *Mind* (106: 263-277).

Reichenbach, H., (1956) *The Direction of Time*. Berkeley: University of California Press.

Reichenbach, H., (1976) *Laws, Modalities, and Counterfactuals*. Berkeley: University of California Press.

Russell, B., (1913) "On the Notion of Cause", *Proceedings of the Aristotelian Society* (13: 1-26).

Ryle, G., (1971) *Collected Papers*, vol. 2, London: Hutchinson.

Salmon, W.C., (1978) "Why ask 'Why'? An Inquiry Concerning Scientific Explanation," *Proceedings and Addresses of the American Philosophical Association* (51: 683-705).

Salmon, W.C., (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, Princeton University Press.

Salmon, W.C., (1985) "Conflicting Conceptions of Scientific Explanation" *Journal of Philosophy* (82: 651-654).

Salmon, W.C., (1990) *Four Decades of Scientific Explanation* Minneapolis: University of Minnesota Press.

Salmon, W.C., (1997) *Causality and Explanation*. Oxford: Oxford University Press.

Savitt, S.F., (ed.) (1995) *Time's Arrows Today: Recent physical and philosophical work on the direction of time*. New York: Cambridge University Press.

Schaffer, J., (2001) "Causes as Probability Raisers of Processes" *Journal of Philosophy* (98: 75-92).

Scheines, R., (1997) "An Introduction to Causal Inference," in V.R. McKim and S. Turner, (eds.) *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, Indiana: University of Notre Dame Press (185-200).

Shafer, G., (1996) *The Art of Causal Conjecture*, Cambridge: MIT Press.

Shafer, G., (1997) "How to think about causality" MS, Faculty of Management, Rutgers University.

Shafer, G., (2000) Comment. *Journal of the American Statistical Association* (95: 438-442).

Shipley, B., (2000) *Cause and Correlation in Biology: A Users Guide to Path Analysis, Structural Equations and Causal Inference*, Cambridge: Cambridge University Press.

Simon, H.A., (1953) "Causal ordering and Identifiability", in W.C. Hood and T.C. Koopmans, (eds.), *Studies in Economic Method*, New York: Wiley, (49-74).

Singpurwalla, N., (2002) "On Causality and Causal Mechanisms (with discussion)" *International Statistical Review* (10: 2: 198-206).

Sklar, L., (2003) "Dappled Theories in a Uniform World", *Philosophy of Science* (70: 2: 424-441).

Skyrms, B., (1980) *Causal Necessity*, New Haven: Yale University Press.

Sosa, E. and M. Tooley, (eds.) (1993) *Causation*, Oxford Readings in Philosophy (ser.), New York: Oxford University Press.

Spirtes, P., (1994) "Conditional Independence in Directed Cyclic Graphical Models for Feedback." Technical Report CMU-PHIL-54, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Spirtes, P., C. Glymour and R. Scheines, (1993) *Causation Prediction and Search*. New York: Springer-Verlag.

Spohn, W., (1980) "Stochastic Independence, Causal Independency, and Shieldability" *Journal of Philosophical Logic* (9: 73-99).

Spohn, W., (2001) "Bayesian Nets Are All There Is to Causality," in *Stochastic Dependence and Causality*. D. Constantini, M.C. Galavotti, and P. Suppes (eds.), Stanford CSLI Publications.

Strawson, P.F., (1959) *Individuals: An Essay in Descriptive Metaphysics*, London: Methuen.

Suppe, F., (ed.) (1977) *The Structure of Scientific Theories*, (Second Edition), University of Illinois Press: Chicago.

Suppe, F., (1989) *The Semantic Conception of Theories and Scientific Realism*, Urbana: University of Illinois Press.

Suppe, F., (2000) "Understanding Scientific Theories: An Assessment of Developments, 1969-1998" *Philosophy of Science* (67: S103-S115).

Suppes, P., (1962) "Models of Data", in E. Nagel, P. Suppes and A. Tarski, (eds.) *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford: Stanford University Press (252-261).

Suppes, P., (1970) *A Probabilistic Theory of Causality*, Amsterdam: North Holland.

Tabery, J.G., (2004) "Synthesizing Activities and Interactions in the Concept of a Mechanism" *Philosophy of Science* (71: 1-15).

Teller, P., (2001) "Twilight of the Perfect Model Model" *Erkenntnis* (55: 393-415).

Thomason, R., (2003) "Logic and Artificial Intelligence," *The Stanford Encyclopaedia of Philosophy* (Fall 2003 Edition), E.N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2003/entries/logic-ai/.

Tooley, M., (1993) "Causation: Reductionism Versus Realism," in E. Sosa, and M. Tooley, (eds.) *Causation*. Oxford Readings in Philosophy (ser.), New York: Oxford University Press, (172-192).

Tooley, M., (2003) "The Stalnaker-Lewis Approach to Counterfactuals" *The Journal of Philosophy* (100: 7: 371-377).

Tversky, A., and D. Kahneman, (1980) "Causal schemata in judgements under uncertainty", in M. Fishbein, (ed.) *Progress in Social Psychology*. Hillsdale, NJ: Erlbaum, (49-92)

Twardy, C.R., and K.B. Korb, (2002) *Causal Interaction in Bayesian Networks*, School of Computer Science and Software Engineering, Monash University, Melbourne, Technical Report #118 (1-11).

Van Fraassen, B.C., (1980) *The Scientific Image*. Oxford: Clarendon Press.

Van Fraassen, B.C., (1987) "The Semantic Approach to Scientific Theories," in N.J. Nersessian, (ed.) *The Process of Science*, Dordrecht: Martinus Nijhoff, (105-124).

Van Fraassen, B.C., (1989) *Laws and Symmetry*, New York: Oxford University Press.

Von Wright, G.H., (1971) *Explanation and Understanding*. New York: Cornell University Press.

Von Wright, G.H., (1973) "On the Logic and Epistemology of the Causal Relation," in P. Suppes, (ed.) *Logic, Methodology and Philosophy of Science IV*, Amsterdam: North Holland.

Waldmann, M.R., K.J. Holyoak, and A. Fratiannea (1995) "Causal Models and the Acquisition of Category Structure," *Journal of Experimental Psychology* (124: 181-206).

Williams, D., (1963) *The Ground of Induction*, New York: Russel and Russel.

Wimsatt, W., (1974) "Complexity and Organization" in K. Schaffner and R. Cohen (eds.), *PSA 1972: Boston Studies in the Philosophy of Science* (20: 67-86).

Winsberg, E., (2003) "Simulated Experiments: Methodology for a Virtual World" *Philosophy of Science* (70: 1: 105-125).

Wittgenstein, L., (1976) "Cause and Effect: Intuitive Awareness," R. Rhees, (ed.) P. Winch (trans.) *Philosophia* (6: 392-445).

Woodward, J., (2000) "Explanation and Invariance in the Special Sciences" *British Journal for the Philosophy of Science* (51: 197-254).

Woodward, J., (2001) "Causation and Manipulability", *The Stanford Encyclopaedia of Philosophy* (Fall Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2001/entries/causation-mani/>

Woodward, J., (2002a) "What is a Mechanism? A Counterfactual Account" *Philosophy of Science* (69: S366-S377).

Woodward, J., (2002b) "Invariance, Modularity, and All That: Cartwright on Causation" paper presented at *Nancy Cartwright's Philosophy of Science: An International Workshop*. University of Konstanz.

Woodward, J., (2003) "Scientific Explanation," The Stanford Encyclopaedia of Philosophy (Summer Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2003/entries/scientificexplantion/>.

Wright, S., (1921) "Correlation and Causation" *Journal of Agricultural Research* (20: 557-585).